International Journal of Management Science
Vol 8, No 2, November 2002

Minimizing Production Lead Time of Kanban System in a Stochatic Environment

Ilhyung Kim[*]

School of Business Administration, Ajou University,
San 5 Wonchon-Dong, Padal-Gu, Suwon, 442-749, South Korea

(Received Sep. 2001, Revised May 2002, Accepted Sep. 2002)

ABSTRACT

This paper presents a model that analyzes the impact of uncertainties in demand and processing times on the production lead time of a Kanban system. We consider the waste associated with under-production as well as over-production when we measure the production lead time. We set up an optimization model to minimize the production lead time. A simple heuristic procedure is developed to determine solutions in terms of the size of containers and the number of Kanban cards. In addition, we numerically examine the behavior of the optimal Kanban system.

1. INTRODUCTION

According to a survey conducted by Miller and Roth [13] and the research work reported in Blackburn [4], lead time has become an important strategic weapon for a manufacturer to compete in the world market. In operations management literature, a significant amount of research has been devoted to analyzing lead time in manufacturing systems. Much of the research has been concentrated on flow time within the scheduling context. In this context, the processing time of a job on a machine is assumed to be known and fixed (see Lenstra et al. [12] for an extensive review of scheduling problems). Other analytical research has focused on cycle time and/or set-up time reduction [15, 19]. Others have considered the problem of setting due-dates [6, 18]. Finally, a number of researchers have considered the relationship between lead times and other parameters in manufacturing systems [2, 3, 9, 10, 11, 16, 21].

[*] Email: ikim@madang.ajou.ac.kr

Traditional inventory models such as MRP and EOQ do not account for the effects of uncertainties on lead time. In most stochastic models, the lead times are assumed to be independent and identically distributed, which do not capture the reality of the most manufacturing systems. Karmarkar [9] developed a model to examine the relationship between lead time and batch size. However, in his model, Karmarkar has ignored the effect of batch size on shortages and finished goods inventories. In addition, no systematic inventory control mechanism such as the time and the amount to release the order to the shop was established. Bitran and Tirupati [3] further developed a more comprehensive model that integrates the job release mechanism into Karmarkar's model. They considered the lotsizing problem with stochastic demands and stochastic lead times under a continuous review policy of (R, Q) type. They presented some characteristics of their problem and suggested several ways to simplify and solve the problem. They also examined the effect of various approximation schemes on different performance measures.

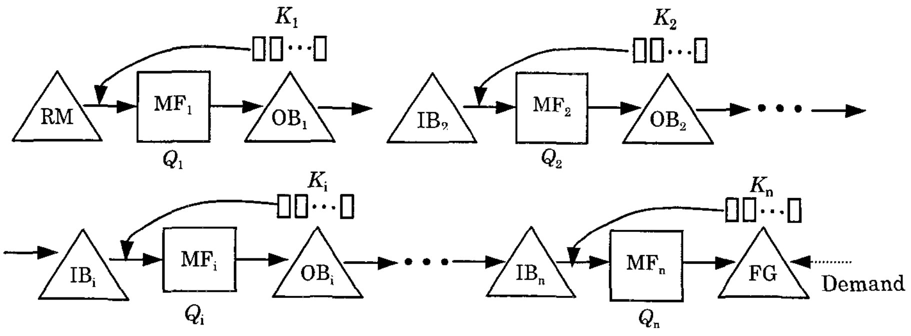
This paper presents a model that analyzes the impact of uncertain demand and uncertain processing time on total production lead time under a Kanban system. The Kanban system is originally designed by Toyota Motor Co. to realize just-in-time production. The competitive and strategic advantages associated with the Kanban system are reducing the cost of inventories, increasing the plant capacity, cutting direct labor costs, improving quality, enhancing flexibility of the production system, and shortening the lead time [7]. The reader is referred to Berkley [1] for an extensive review of Kanban production control literature.

Another important aspect of our model is the way we treat the production lead time. Traditionally, the queuing and processing time at a manufacturing facility has been considered as a production lead time. As Ohno [14] pointed out, not only the waiting time of items at the inventory buffer (waste of over-production) but also the waiting time of orders from a succeeding operation (waste of under-production) clearly ruin the efficiency of the system. These times can be substantial and may consist of over 50% of total time in the system especially in a job shop environment. This point has been neglected in the literature, and motivated us to develop a model in which the waste associated with under-production as well as over-production is incorporated into total production lead time. In order to compute that wastes, we conceptually divide the inventory buffer into two types: *in*-and *out-buffer* before and after each manufacturing facility. The detailed role of each buffer is explained as follows: when job is released to the facility, it is sent to the in-buffer. If the facility is idle, then the job released to the facility is immediately processed, otherwise it has to wait until the facility is available. After being processed, the completed item is sent to the out-buffer. If the demand is backordered at the out-buffer, then the item is immediately used to fulfill the demand backordered, otherwise it has to wait at the out-buffer until the next demand arrives. Therefore, the in-buffer captures the traditional production lead time such as queueing and processing time at the manufacturing facility, and the out-buffer captures the times pointed out by Ohno such as the waiting time at the inventory buffer and the waiting time of demand backordered.

By considering a Kanban control system, we develop expressions for measuring the total production lead time. We establish a simple heuristic procedure to determine solutions for the size of the containers and the number of Kanban cards. In addition, the behavior of optimal Kanban system is investigated by means of numerical experiments. The remainder of the paper is organized as follows: The model is presented in Section 2, and the characteristics of performance measures and the searching procedure to determine the optimal policy is developed in Section 3. The results of computational experiments are discussed in Section 4. Finally, we conclude the paper with some future extensions.

2. THE MODEL

As depicted in Figure 1, the system consists of a series of processors and two inventory buffers - in-buffer and out-buffer - for each processor. We assume, in this paper, that production and inventory is controlled by a Kanban system, and the problem is to determine the size of the containers and the number of Kanban cards for each operation so that the total production lead time is minimized.



MF_i : manufacturing facility i
RM : raw materials inventory (in-buffer of MF_1)
IB_i : in-buffer of MF_i
OB_i : out-buffer of MF_i
FG : finished goods inventory (out-buffer of MF_n)
K_i : number of Kanban cards at MF_I
Q_i : batch size of MF_i

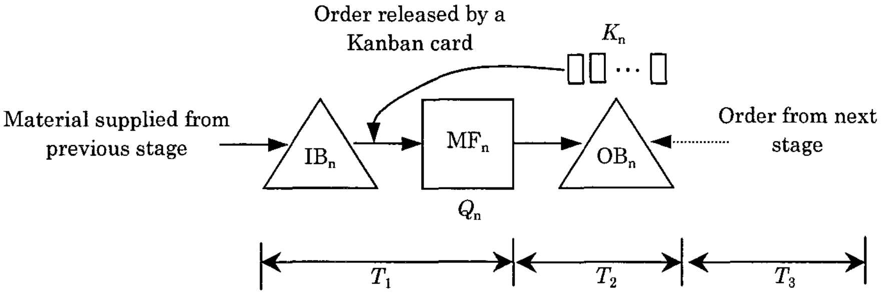
Figure 1. Configuration of Serial Production System

At each stage the completed items are stored in containers, each of which holds exactly Q_i units. There are K_i containers at stage i, and there is a Kanban card attached to each container. The Kanban system operates in the following way: Whenever Q_i units are depleted from a container in the ship-buffer i, the corresponding Kanban card is transmitted to the in-buffer i, where it withdraws raw materials from the preceding operation (out-buffer i-1). The Kanban cards also serve as new production orders that trigger the processor to begin its production process. In general, the processor uses a first-come-first-serve discipline to process these orders. Once the processor i produces Q_i units, the completed units and the Kanban card which ordered the full container are sent to the out-buffer i. In the event when the succeeding operation (stage i +1) places an order and there is no completed items available in the ship-buffer i, this order should wait until the completed items become available. In our model, we assume that the set-up time is incurred when the processors begin their production for each (container) of their orders. One can think of processor as an oven, a chemical tank, or a dyeing machine.

There are four important observations regarding the Kanban system described above. First, the number of Kanban cards circulating at stage i is K_i at any point in time. Hence, the maximum inventory level in out-buffer i is equal to K_i Q_i . Second, a Kanban card is sent to in-buffer i (an order is released) whenever a full container (Q_i units) is depleted at the out-buffer i. Therefore, the number of Kanban cards, K_i , and the container size, Q_i , dictate the arrival process of the Kanban cards at the in-buffer. Third, the container size (or batch size), Q_i , affects the intensity of stage i. Smaller batches cause the workload on the processor to increase due to increased number of set-ups. As batches become large, the intensity of the stage decreases and the effect of set-up times diminishes. Finally, the number of Kanban cards, K_i , affects the inventory level and the number of backorders at the out-buffer i. For instance, as the number of Kanban cards increases, the expected inventory increases while the expected number of backorders decreases.

This operating characteristics of the system can be analyzed by a serial queuing system. However, it is difficult to analyze the system exactly. In order to make the model tractable, we assume that the system can be decomposed into independent queuing systems, each of which has a single processor and two buffers before and after the processor [20]. As an initial study of this decomposition scheme, we shall restrict our analysis to a single stage queuing system.

Consider a single-stage system as shown in Figure 2. In this system, we can convert the units of goods in terms of number of containers (or number of Kanban cards). Thus, at any given point in time, the state of the system can be specified by (1) the number of Kanban cards waiting at the in-buffer be processed and the number of Kanban cards being processed at the processor ; (2) the number of Kanban cards in the out-buffer ; and (3) the number of backorders. Let , T_1 be the lead time that consists of the average time that a Kanban card waits to be processed at the in-buffer and the average processing time, T_2 be the average time that a Kanban card waits at the out-buffer, and T_3 denote the average time that an order from the succeeding operation waits for completed items.



T_1 : queuing time at the in-buffer and processing time
T_2 : waiting time at the out-buffer
T_3 : waiting time of the order from the next stage

Figure 2. Schematic Representation of Production Lead Time for a Decomposed Single Facility

In order to determine an efficient Kanban system (specified by K, the number of Kanban cards, and Q, the size of the container) that minimizes total production lead time, we formulate the following mathematical program:

$$(P1) \qquad \min_{Q,K} T = T_1 + T_2 + T_3 \qquad (1)$$

We make the following assumptions for the analysis:

- The arrival process of each unit demand (order from the succeeding operation) is a Poisson process. In our model, a Kanban card is transmitted to the in-buffer after every Q units are depleted from a container in the out-buffer. Thus, the interarrival times of the Kanban cards at the in-buffer have an Erlang distribution of order Q.
- All unmet demands are backlogged.
- The sum of the time needed to produce Q units (a batch) and the set-up time is exponentially distributed. While it is difficult to justify use of an exponential distribution for production time, this assumption is made to keep the model tractable [3, 9]. However, if both the set-up time and the processing time are

highly uncertain, this assumption can be considered to be reasonable. Exponential production times seem to be reasonable for modeling semiconductor manufacturing processes [5].

We now turn our attention to developing the expressions for T_1, T_2, and T_3. To do so, let us define

D : demand rate of the product (induced from next facility)
Q : batch size
K : number of Kanban cards
λ : arrival rate of batches at the machine = D / Q
τ : set-up time for a batch
P : unit production rate of machine
$\bar{x}$: production time per batch = $\tau + (Q / P)$
μ : processing rate at the machine $= 1/\bar{x} = P/(P\tau + Q)$
$\bar{I}$: expected on-hand inventory
$\overline{B}$: expected number of backorders

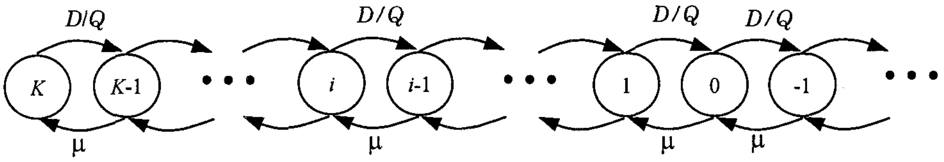
Since the Kanban card that triggers production arrives the in-buffer after Q jobs have been depleted at the out-buffer, the job inter-arrival times to the facility have an Erlang distribution of order Q and mean Q / D. The number of job at the machine is bounded by K. Therefore T_1 requires computation of steady state probabilities of an $E_Q / M / 1 / K$ system. In evaluating this measure, the job arrivals to the shop are approximated by a Poisson process with the parameter D/Q. Using this approximation, T_1 is determined by $M/M/1/K$ system that can be expressed as [17]:

$$T_1 = \frac{(1 + K(D\tau/Q + D/P)^{K+1} - (K+1)\,(D\tau/Q + D/P)^K)\,(\tau + Q/P)}{(1 - D\tau/Q - D/P)\,(1 - (D\tau/Q + D/P)^{\,K})} \tag{2}$$

The on-hand inventory and the number of backorders at the out-buffer can be described by a continuous time Markov chain (Figure 3). The state i corresponds to the case when the number of Kanban cards in the inventory buffer is i. Let $P_{K\text{-}n}$ be the limiting probability that a system will be in state K–n. The standard results from probability models give [17]

$$P_{K-n} = (D/P + D\tau/Q)^n (1 - D/P - D\tau/Q) \tag{3}$$

When the system is in state K–n, the probability that inventory position is x is $1/Q$, for $(K-n-1)Q < x \le (K-n)Q$. The expected on-hand inventory at the out-buffer is given by



D/Q : inventory depletion rate from succeeding facility
μ : arrival rate of Kanban cards to inventory buffer

Figure 3. State Transition Diagram of Kanban Cards

$$\bar{I} = \sum_{n=0}^{K-1} \big((K-n)Q + Q/2\big) P_{K-n} \tag{4}$$

The average waiting time of a job at the out-buffer is determined by dividing the expected on-hand inventory by demand rate as follows:

$$T_2 = \bar{I}/D = \frac{Q}{D}\left(K - \frac{1}{2} - \frac{(D\tau/Q + D/P)(1-0.5(D\tau/Q+D/P)^K) - 0.5(D\tau/Q+D/P)^{K-1}}{(1-D\tau/Q-D/P)}\right) \tag{5}$$

The expected number of backorders is given by

$$\bar{B} = \sum_{n=K}^{\infty} \big((n-K)Q + Q/2\big) P_{K-n} \tag{6}$$

The average time that an order from the succeeding operation waits for completed items is determined by dividing the expected backorders by demand rate as follows:

$$T_3 = \bar{B}/D = \frac{Q(D\tau/Q + D/P)^K (1 + D\tau/Q + D/P)}{2D(1 - D\tau/Q - D/P)} \tag{7}$$

Using Poisson approximation scheme, the problem is finally formulated as follows:

$$(P2) \qquad \min_{Q,K} \; T = T_1 + T_2 + T_3 \tag{8}$$

$$\text{s.t.} \quad K \in I^+ \tag{9}$$

$$Q \geq D\tau/(1 - D/P) \tag{10}$$

where, T_1, T_2 and T_3 are as in equations (2), (5) and (7) respectively. Since we must have system intensity, $\rho = \lambda/\mu < 1$, the batch size, Q, is bounded below by $D\tau/(1-D/P)$ which is shown in constraint (11).

3. CHARACTERISTICS OF OBJECTIVE FUNCTION AND SEARCHING PROCEDURE

To determine optimal values of Q and K so that the total production lead time, T, is minimized, we develop a search procedure. As shown in equations (2), (5) and (7), the total production lead time, T, is a very complicated function of Q and K. It is difficult to get a mathematically-nice and completely-proven properties for T. In order to develop alternative solution approaches, we analyze the characteristics of T_1, T_2, and T_3 that are summarized in the following lemma and conjectures.

Lemma 1. For a given batch size Q,

1. T_1 is concave and monotonically increasing in K.
2. T_2 is convex and monotonically increasing in K.
3. T_3 is convex and monotonically decreasing in K.

Proof: The proof directly from the following properties:

$\partial T_1 / \partial K > 0, \partial T_2 / \partial K > 0, \partial T_3 / \partial K < 0, \partial^2 T_1 / \partial K^2 < 0, \partial^2 T_2 / \partial K^2 > 0,$ and $\partial^2 T_3 / \partial K^2 < 0.$

The results of Lemma 1 are fairly intuitive. As the number of Kanban cards (K) increases, more jobs are triggered to the manufacturing facility, which in turn increases the queuing and processing time at the in-buffer (T_1). However, since the number of jobs triggered to the facility is constrained by not only the number of Kanban cards but also demand rate, further increase in K may not have significant effect on T_1 when K is sufficiently large. Finally T_1 converges to the case of incapacitated queuing system, $M/M/1/\infty$ as follows:

$$\lim_{K\to\infty} T_1 = (\tau + Q/P)/(1 - D\tau/Q - D/P) = 1/(\mu - \lambda) \tag{11}$$

In addition, increasing the number of Kanban cards causes the expected on-hand inventory at the out-buffer ($\bar{I}$) to increase, which results in increase of the waiting time at the out-buffer (T_2). On the other hand, increasing the number of Kanban cards causes the expected number of backorders to decrease, which results in decrease of the time for the next facility to wait (T_3).

Conjecture 1. For a given Q, T is a bowl-shaped unimodal function in K and has unique minimum point K^* such that $T(K^*) \le T(K), \forall\ K$.

It follows from Lemma 1 that T is the sum of monotonically increasing and monotonically decreasing function. This observation leads us to conjecture that T

is a unimodal function. For a given Q, the second derivative of T with respect to K is given by

$$\frac{\partial^2 T}{\partial K^2}=\frac{Qa^K\ln a}{D(1-a)(1-a^K)^3}\left(-2a(1-a)\left(1-a^K\right)+(1+a)\left(1-a^K\right)^3\ln a-Ka(1-a)\left(1-a^K\right)\ln a\right) \tag{12}$$

where, $a=D/P+D\tau/Q$. For a large K, a^K is close to 0 since $0 < a < 1$, and $1-a^K$ and $1+a^K$ can be approximated to 1. Using these approximation, we can see that

$$\frac{\partial^2 T}{\partial K^2}\begin{cases} >0 & \text{if } K<K^f \\ =0 & \text{if } K=K^f \\ <0 & \text{otherwise} \end{cases} \tag{13}$$

where, $K^f=((1+a)\ln a-2a(1-a))/(a(1-a)\ln a)$. Therefore, T is convex-concave function in K for a given Q. In addition, following property also holds:

$$\lim_{K\to\infty}\frac{\partial T}{\partial K}=\frac{Q}{D}>0 \tag{14}$$

The conditions of equations (13) and (14) give that T is a bowl-shaped unimodal function in K for a given Q.

Conjecture 2. For a given K, T is a bowl-shaped unimodal function in Q and has unique minimum point Q^* such that $T(Q^*) \le T(Q)$, $\forall$ Q.

Let us consider T_1 and T_3 simultaneously. Notice that the number of jobs in manufacturing facility is bounded by the number of Kanban cards. If the number of Kanban cards is fairly large, no demand from the next facility is waiting at the out-buffer, which makes T_3 virtually zero. As we decrease the number of Kanban cards, some of the queuing time (T_1) at the in-buffer may be shifted to the waiting time of demand from next facility (T_3). Therefore, the sum of T_1 and T_3 can be approximated by the total queuing and processing time of $M/M/1/\infty$ system, which has been shown to be convex in Q, as discussed in Karmarkar [9]. On the other hand, Kanban system with K Kanban cards can be viewed as a special case of the (R, Q) policy, where R = (K–1)Q. Hence, the average inventories can be decomposed approximately into cycle stock and safety stock. Therefore, T_2 can be reduced to a linear function in Q as follows:

$$\bar{I}=\text{cycle stock}+\text{safety stock}=Q/2+(K-1)\,Q \tag{15}$$

Finally, we conjecture that T is a sum of convex (T_1 and T_3) and linear(T_2) function in Q, which implies that T is a bowl-shaped unimodal function in Q for a given K.

Conjecture 3. Let $Q^*(K)$ be an optimal batch size for a given K. Then $Q^*(K)$ is decreasing in K.

As discussed in Bitran and Tirupati [3], the total production lead time using M/M/1 queue and the standard inventory approximation under a continuous review policy of (R, Q) type can be formulated as follows:

$$T = \frac{Q\rho}{D\,(1-\rho)} + \frac{2B(Q,R)}{D} + \frac{1}{D}\left(R + \frac{Q+1}{2} - \frac{\rho}{1+\rho} \right) \tag{16}$$

where,

$$B(Q,R) = \text{expected number of backorders} = \frac{(1-\rho^Q)\rho^{R+2}}{Q(1-\rho)^2}$$

$$\rho = \text{intensity of the system} = D/P + D\tau/Q$$

$$R = \text{reorder point}$$

Suppose we differentiate T with respect to R. The first order optimality condition can be expressed as:

$$\rho^{R^*+2} = \frac{-(1-\rho)^2 Q}{2\,(\ln\rho)(1-\rho^Q)} \tag{17}$$

Since left hand side of equation (17) is decreasing in R^* and right hand side is increasing in Q, optimal reorder point level for a given Q, $R^*(Q)$, is decreasing in Q. It is not uncommon to conjecture that the reverse relationship is also true; $Q^*(K)$ is decreasing in K.

Based on above conjectures, we develop a simple searching procedure to determine optimal batch size (Q^*) and optimal number of Kanban cards (K^*). Furthermore, *Conjecture 3 allows us to calculate maximum batch size, which occurs* at $K = 1$. T_1, T_2 and T_3 evaluated at $K = 1$ are as follows:

$$T_{1,K=1} = \tau + Q/P \tag{18}$$

$$T_{2,K=1} = \frac{Q(1 - D/P - D\tau/Q)}{2D} \tag{19}$$

$$T_{3,K=1} = \frac{Q(D/P + D\tau/Q)(1 + D/P + D\tau/Q)}{2D(1 - D/P - D\tau/Q)} \tag{20}$$

First order condition for equations (18), (19) and (20) gives

$$Q_{\max} = \frac{D\tau}{1-D/P}\left(1+\sqrt{\frac{2}{1+D/P}}\right) \tag{21}$$

As D/P approaches to 1, which represents very high traffic density, $Q_{\max}$ approaches 2 $Q_{\min}$, where $Q_{\min}$ is derived from the condition that $\rho = \lambda / \mu < 1$ (Equation 10).

Since we have upper and lower bound for Q and T is a bowl-shaped unimodal function in Q for a given K, $Q^*(K)$ can be easily determined by a binary search. In addition, since T is also unimodal in K for a given Q, $K^*(Q)$ can be easily determined by a usual numerical approximation. Finally, the solution can be determined by the following iterative procedure, which is called the Iterative Searching Heuristic (ISH):

Step 0. Find $Q_{\min}$ and $Q_{\max}$. Let Q^*(K) = $Q_{\max}$.
Step 1. Let $Q = Q^*(K)$. Starting from $K = 1$, increase K by 1 until find $K^*(Q)$.
Step 2. Let $K = K^*(Q)$. Find $Q^*(K)$ by using binary search for ($Q_{\min}$, $Q_{\max}$).
Step 3. If $Q^*(K)$ and $K^*(Q)$ remain same, then stop, $Q^* = Q^*(K)$, and $K^* = K^*(Q)$. Otherwise go to Step 1.

In ISH, conjecture 2 is not necessarily needed. Without this conjecture, we need a complete search between lower and upper bound, which increases the number of steps in computation.

4. COMPUTATIONAL EXPERIMENTS AND SIMULATION

The objective of the computational experiments is two-fold. The first is to verify conjectures made earlier and the second is to investigate the behavior of optimal solution and its relationship to other parameters. In addition, simulation is used to examine the accuracy of Poisson approximation scheme used in Section 2.

In our computational experiments, we design different type of problems by varying the ratio of demand rate to processing rate (D/P) and the ratio of set-up time to processing time (τP). 19 different cases are generated by varying D/P from 0.05 to 0.95 incremented by 0.05, and 29 different cases are generated by varying τP from 0 to 20 incremented by 1 and from 30 to 100 incremented by 10. Total 551 (19 times 29) cases are generated and analyzed for computational experiments. Among these, 9 cases are selected for simulation as follows: $D/P = 0.4$, 0.6 and 0.8 which represent low, medium and high traffic density respectively, and $\tau P = 1$, 5 and 10 which represent short, medium and long set-up time respectively.

In all 551 cases of computational experiments, all three conjectures discussed in Section 3 are valid. Examples of computational experiments for Conjectures 1, 2, and 3 are illustrated in Figures 4, 5, and 6, respectively. As shown in the figures, numerical experiments clearly verify all three conjectures. Figure 6 also shows the results of simulation, which generally support Conjecture 3. However, when the traffic intensity of the system is high ($D/P = 0.8$, SIM), the relationship between the optimal batch size and the number of Kanban cards is not clear. It is suggested to do more or longer simulation runs to get more stable results.

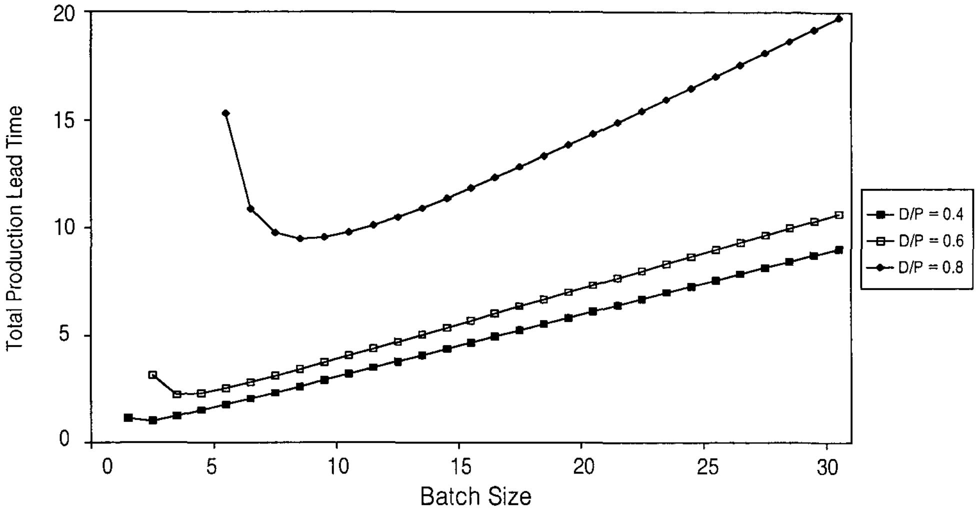


Figure 4. Total Production Time vs. Batch Size (when $K = 1$, $\tau P = 1$)

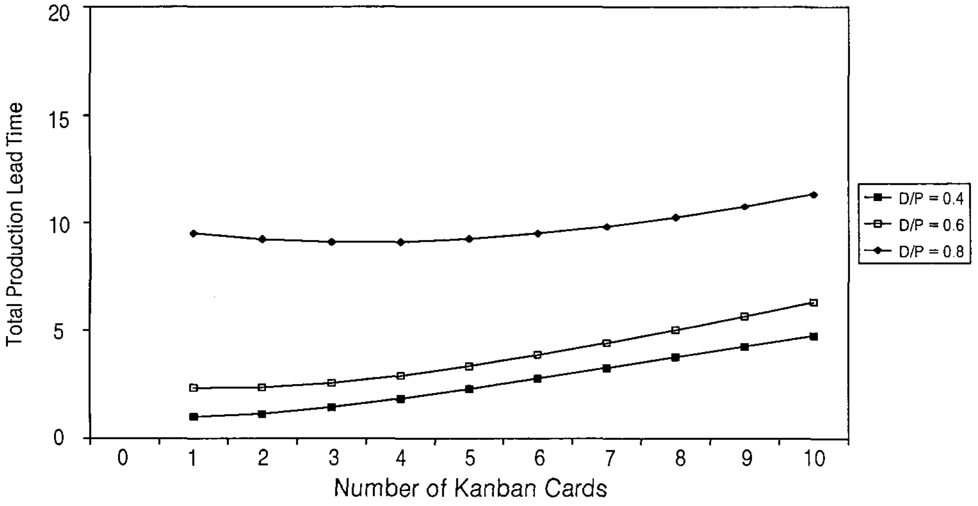


Figure 5. Total Production Time vs. Number of Kanban Cards ($Q = Q^*$, $\tau P = 1$)

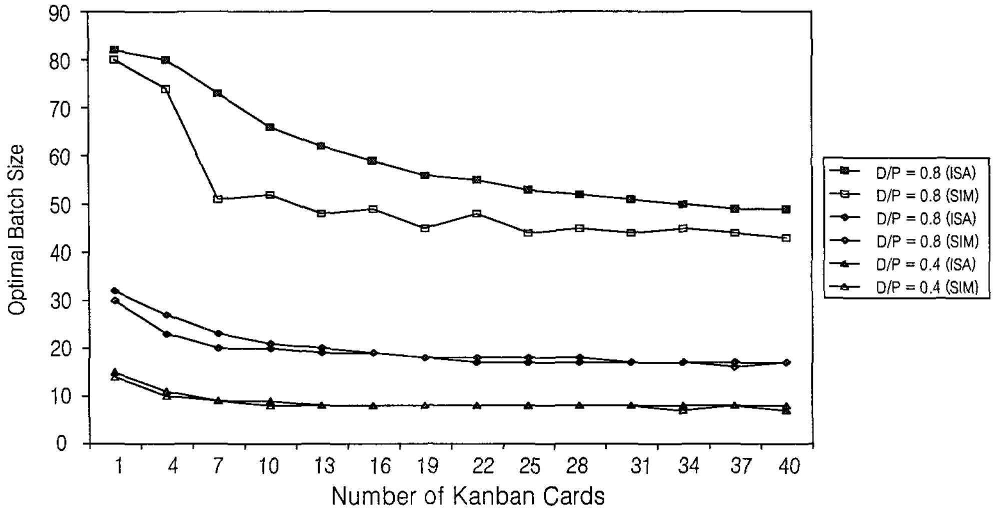


Figure 6. Optimal Batch Size vs. Number of Kanban Cards (when $\tau P = 10$)

Table 1 compares the production lead time calculations of ISH and simulation for 9 cases. It is apparent from the table that use of Poisson approximation underestimates the queuing and processing time at the in-buffer of manufacturing facility (T_1), and seriously overestimates the waiting time of demand (T_3) especially for the case of high utilization rate (*D*/*P*).

Table 1. Lead Time Calculations by Heuristic and Simulation

		T_1		T_2		T_3		T	
D /*P*	τP	ISH[a]	SIM[b]	ISH	SIM	ISH	SIM	ISH	SIM
0.4	1	0.300	0.300	0.100	0.137	0.600	0.401	1.000	0.838
0.4	5	1.300	1.301	0.350	0.308	3.604	1.917	4.714	3.526
0.4	10	2.500	2.498	0.625	0.503	6.250	3.887	9.375	6.888
0.6	1	0.500	0.501	0.083	0.085	1.750	1.150	2.233	1.736
0.6	5	2.899	3.306	0.950	1.090	7.200	3.367	11.039	7.763
0.6	10	5.914	6.660	2.023	2.315	14.133	6.079	22.070	15.054
0.8	1	1.737	2.272	0.426	0.554	6.926	3.438	9.089	6.264
0.8	5	8.869	10.967	2.231	3.254	34.323	13.427	45.423	27.648
0.8	10	17.738	21.769	4.462	6.551	68.646	23.379	90.846	51.699

Note: a: Results of iterative searching heuristic
b: Results of simulation using Q^* and K^* determined by ISH

Table 2 shows the comparison of solutions determined by ISH with those determined by simulation. In most cases, the differences of the solutions for both batch size and the number of Kanban cards between the two methods fall within ±1 except the case of high traffic density (D/P = 0.8) and long set-up time (τP = 10). It is also observed that Poisson approximation tends to overestimate both the optimal batch size and the optimal number of Kanban cards, but more extensive experiments should be performed before any firm conclusion is drawn for this observation.

Table 2. Optimal Solutions by Heuristic and Simulation

		Q^*			K^*			T^*		
D/P	τP	ISH	SIM	D[a]	ISH	SIM	D	ISH[b]	SIM	Error[c]
0.4	1	2	2	0	1	1	0	0.839	0.839	0
0.4	5	8	7	1	1	1	0	3.525	3.481	1.26
0.4	10	15	14	1	1	1	0	6.888	6.779	1.61
0.6	1	4	3	1	1	1	0	1.735	1.658	4.64
0.6	5	15	16	-1	2	1	1	7.763	7.502	3.48
0.6	10	31	30	1	2	1	1	15.055	14.682	2.54
0.8	1	8	8	0	3	2	1	6.264	5.564	12.58
0.8	5	41	42	-1	3	2	1	27.648	23.933	15.52
0.8	10	82	74	8	3	2	1	51.699	49.429	4.59

Note: a: D = ISH – SIM
b: Simulated results using Q^* and K^* determined by ISH
c: Error (%) = (T^* of ISH – T^* of SIM)/ T^* of SIM

Now, we discuss about the behavior of optimal solutions and their relationship with other parameters. Figure 4 shows the behavior of total production lead time with batch size analyzed by ISH. As shown in figure, total production lead time has a sharp minimum, and is sensitive to the choice of batch size. This result is quite different from conventional EOQ type analysis, which supports the robustness of batch size decision. Figure 5 represents the same kind of behavior with the number of Kanban cards in the system. This observation is somewhat different from previous one. Total production lead time has a flat minimum, and is relatively insensitive to the choice of number of Kanban cards. It can be concluded from above observations that we have to be more careful to determine batch size rather than to choose the number of Kanban cards in the system.

Figures 7 and 8 illustrate the relationship of optimal batch size (Q^*) with traffic density and set-up time ratio respectively. It is distinct that Q^* occurs near

Q^{max} in most of the cases. Although it is difficult to get a closed form solution for Q^*, we can get an approximate form as follows based on above observations:

$$Q^* \approx Q_{max} = \frac{D\tau}{1-D/P}\left(1+\sqrt{\frac{2}{1+D/P}}\right) \tag{22}$$

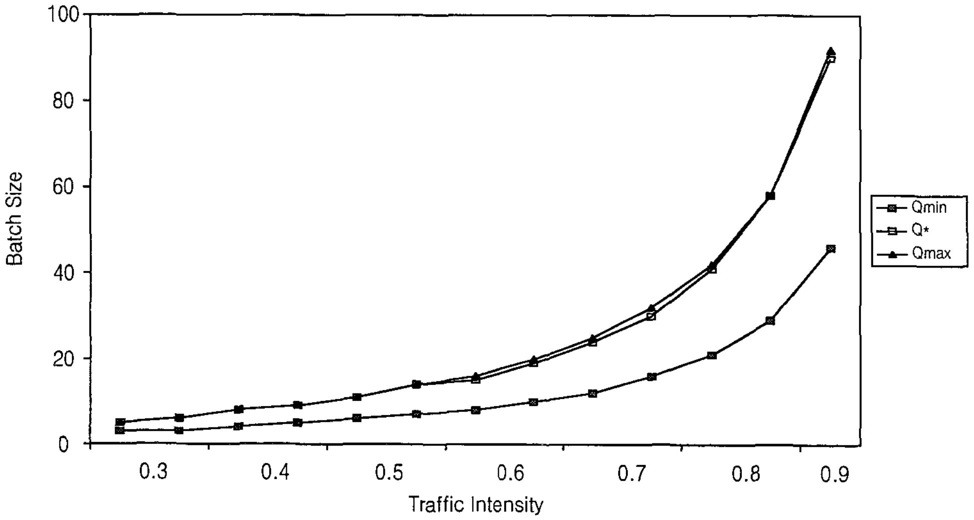


Figure 7. Optimal Batch Size (Q^*), Q_{max} and Q_{min} vs. Traffic Intensity (when $\tau P = 5$)

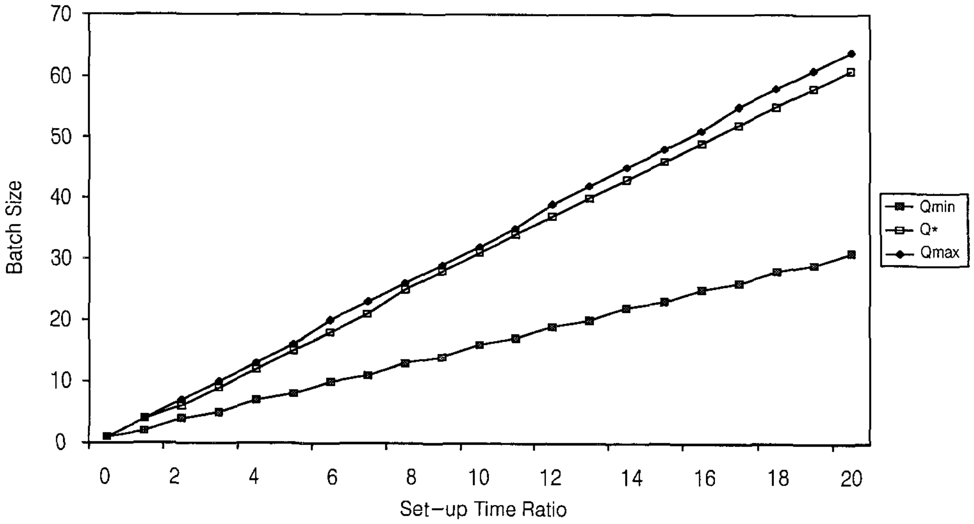


Figure 8. Optimal Batch Size (Q^*), Q_{max} and Q_{min} vs. Set-up Time Ratio (when $D/P = 0.6$)

Notice that Q^* has approximately a linear relationship with the set-up time as shown in equation (22). This observation is also supported by experiments and simulation. In addition, T^* shows a similar pattern (Figures 9 and 10). These results imply that if we reduce set-up time by a certain rate, we can get a reduction in batch size and total production lead time by same rate. As an example, 10% reduction in set-up time would results in 10% reduction in batch size and

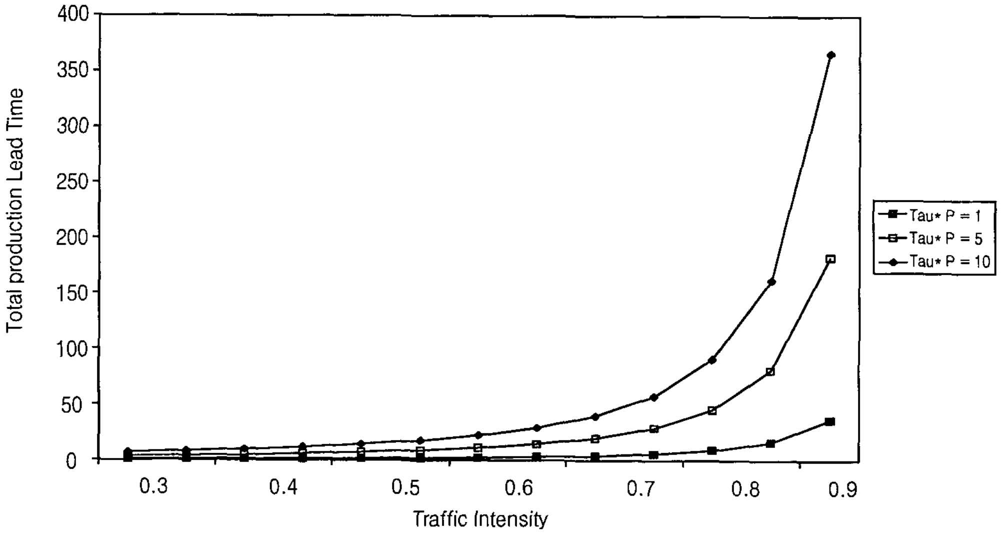


Figure 9. Total Production Lead Time vs. Traffic Intensity

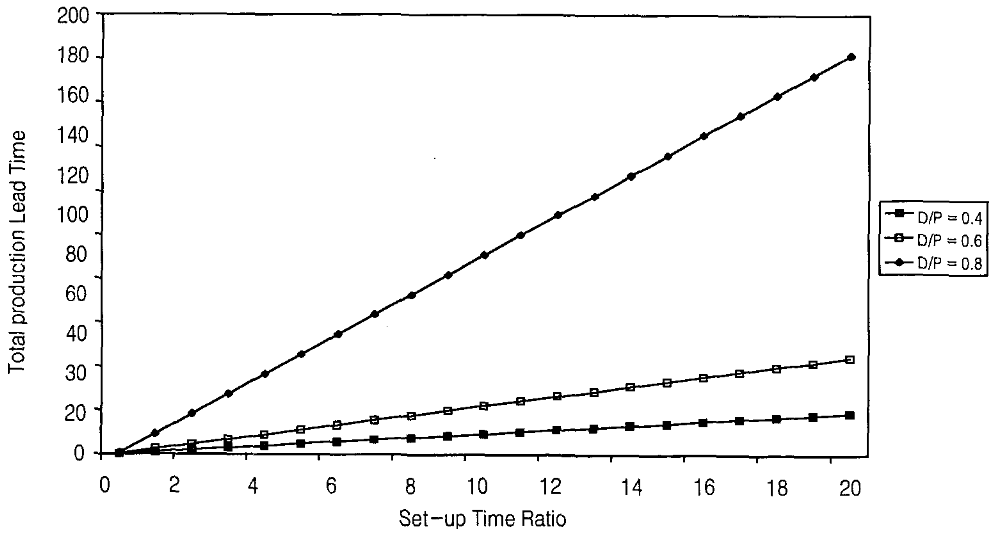


Figure 10. Total Production Lead Time vs. Set-up Time Ratio

total production lead time. It can be also seen that the optimal batch size rises with the traffic density, and is become more sensitive at higher traffic density levels.

Figures 11 and 12 show the relationship of optimal number of Kanban cards (K^*) with different parameters; traffic density and set-up time ratio. It is observed that K^* is virtually independent with set-up time ratio, which can be explained by the following reason. In our system, optimal number of Kanban cards is strongly influenced by the waiting time of demand (T_3), which corresponds to the case of reorder point level in (R, Q) policy. The expected number of backorders which is analogous to T_3 in our system, does not depend on batch size, but only depends on reorder point level in a continuous review system of (R, Q) type [8]. The effect of set-up time changes on system intensity (ρ) can be absorbed by appropriate changes in batch size, since the increase (decrease) in set-up time can be compensated by increase (decrease) in batch size, which makes ρ remain constant ($\rho = D/P + D\tau/Q$). Thus, T_3 can hardly influenced by set-up time changes except the case of very high traffic density.

As shown in Figure 11, optimal number of Kanban cards can be divided into three regions according to traffic density ; first, only one Kanban card is needed for low level of traffic density ($D/P \leq 0.6$), second, two or three Kanban cards are enough for medium level ($0.6 \leq D/P \leq 0.8$), and finally more than three Kanban

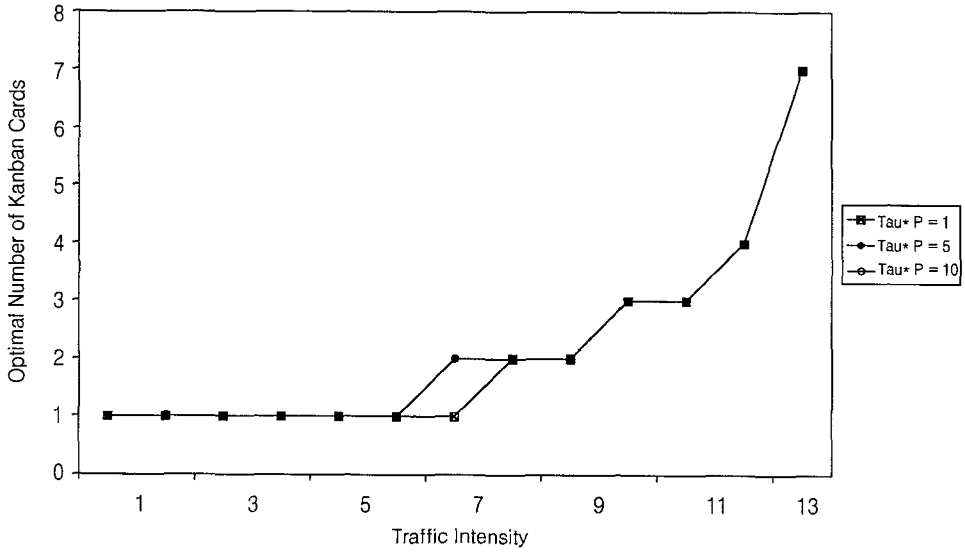


Figure 11. Optimal Number of Kanban Cards vs. Traffic Intensity

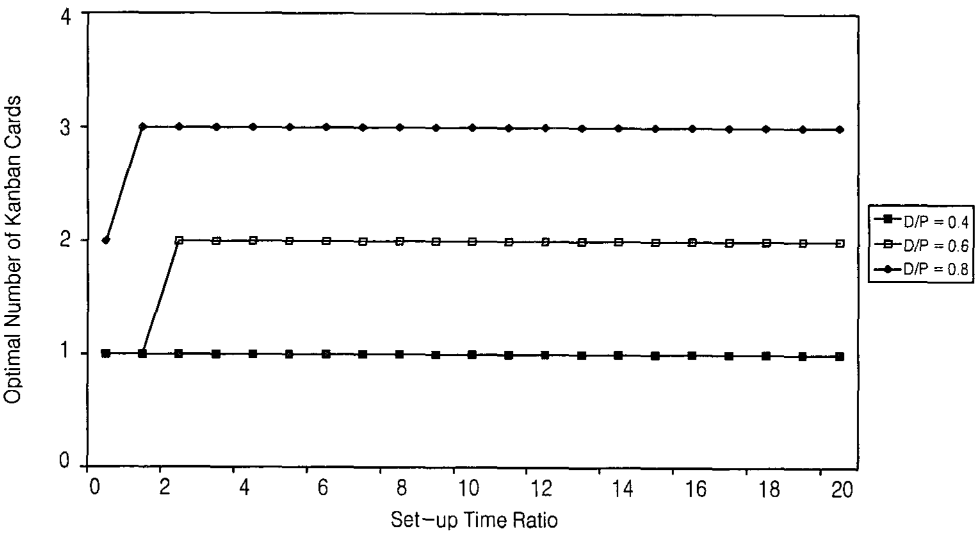


Figure 12. Optimal Number of Kanban Cards vs. Set-up Time Ratio

cards are needed for high level ($D/P \geq 0.8$). Since more Kanban cards make the system more difficult to control and total production lead time is somewhat insensitive to the number of Kanban cards as discussed earlier, it can be implied that at most three Kanban cards are enough for high level of traffic density.

In summary, we suggest, for practical application, simple rules of Kanban control system as follows:

Rule 1. Operate the system with a batch size defined by $Q_{\max}$.

Rule 2. If traffic density is low, just use one Kanban card. As traffic density increases, add one Kanban card at a time. However, there is no need to consider more than 3 Kanban cards.

Rule 3. Use linear relationship between set-up time and total production lead time to evaluate the value of set-up time reduction.

5. CONCLUSIONS

In this paper, we have developed a model which incorporates production and inventory control policy simultaneously. By considering a Kanban control system, expressions were developed for measuring total production lead time. Poisson approximation scheme was used for Erlang distribution. In addition, a heuristic (iterative searching heuristic) was established to determine a solution (batch size and the number of Kanban cards) which minimizes total production lead time. Although total production lead times determined by this heuristic differ substantially from the results analyzed by simulation, it appears that approximation scheme and searching procedures developed in this paper may provide reasonably good solutions. The behavior of optimal solutions and their relationship with other parameters were examined through extensive computational experiments. While the total production lead time has a sharp minimum and is sensitive to the choice of batch size (Figure 4), it has a somewhat flat minimum and is relatively insensitive to the choice of the number of Kanban cards (Figure 5). Finally, we recommended some simple rules for practical application of Kanban system.

Even after we decomposed a series of manufacturing facilities into independent single facility, it was still intractable if we use the direct computational methods. In order to analyze the inter-related effect of preceding and succeeding facilities, it is needed to suggest reasonable simplification and/or useful qualitative properties. This model can be viewed as a starting point of decomposition scheme that has a potential to be extended to more complex situations such as multi-item, multi-facility system.

REFERENCES

[1] Berkley, B. J., "A Review of the Kanban Production Control Research Literature," *Production and Operations Management* 1, 4 (1992), 393-411.

[2] Bertrand, J. W. M., "Multiproduct Optimal Batch Sizes with In-Process Inventories and Multi Work Centers," *IIE Transactions* 17, 2 (1985), 157-163.

[3] Bitran, G. R. and D. Tirupati, "Lot Sizing Under (Q,R) Policy in a Capacity Constrained Manufacturing Facility," *Robotics & Computer-Integrated Manufacturing* 1, 3/4 (1984), 327-337.

[4] Blackburn, J., *Time Based Competition,* Richard D. Irwin, Homewood, Illinois, 1991.

[5] Chen, H., M. J. Harrison, A. Mandelbaum, A. VanAckere, and L. M. Wein, "Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication," *Operations Research* 36, 2 (1988), 202-215.

[6] Duenyas, I. and W. J. Hopp, "Quoting Customer Lead Times," *Management Science* 41, 1 (1995), 43-57.

[7] Henderson, B. D., "The Logic of Kanban," *The Journal of Business Strategy,* 6, 3 (1986), 6-12.

[8] Johnson, L. A., and D. C. Montgomery, *Operations Research in Production Planning, Scheduling, and Inventory Control*, John Wiley & Sons, New York, 1974.

[9] Karmarkar, U. S., "Lotsizes, Lead Times and In-Process Inventories," *Management Science* 33, 3 (1987), 409-418.

[10] Karmarkar, U. S., S. Kekre, and S. Kekre, "The Dynamic Lot-Sizing Problem with Startup and Reservation Costs," *Operations Research* 35, 3 (1987), 389-398.

[11] Kim, I and C. Tang, "Lead Time and Response Time in a Pull Production Control System," *European Journal of Operational Research* 101 (1997), 474-485.

[12] Lenstra, J. K., A. Rinnooy Kan, and P. Brucker, "Complexity of Machine Scheduling problems," *Annals of Discrete Mathematics* 1 (1977), 343-362,.

[13] Miller, J., and A. Roth, "Executive Summary of the 1988 North American Manufacturing Futures Survey," *Boston University Roundtable, Manufacturing*, 1988.

[14] Ohno, T., *Toyota Production System: beyond Large-Scale production*, Productivity Press, Cambridge, 1988.

[15] Porteus, E. V., "Optimal Lot Sizing, Process Quality Improvement and Setup Cost Reduction," *Operations Research* 34, 1 (1986), 137-144.

[16] Rees, L. P., P. R. Philipoom, B. W. Taylor, and P. Y. Huang, "Dynamically Adjusting the Number of Kanbans in a Just-in-Time production System Using Estimated Values of Lead Time," *IIE Transactions* 19, 2 (1987), 199-207.

[17] Ross, S. M., *Introduction to Probability Models*, Academic Press, San Diego, 1989.

[18] Seidmann, A. and M. Smith, "Due Date Assignment for Production Systems," *Management Science* 27, 4 (1981), 401-413.

[19] Spence, A. M. and E. V. Porteus, "Setup Reduction and Increased Effective Capacity," *Management Science* 33, 10 (1987), 1291-1301.

[20] So, K. C. and S. C. Pinault, "Allocating Buffer Storages in a Pull System," *International Journal of Production Research* 26, 12 (1988), 1959-1980.

[21] Zipkin, P. H., "Models for Design and Control of Stochastic Multi-Item Batch Production Systems," *Operations Research* 34, 1 (1986), 91-104.