INPUT CONTROL FOR SYSTEMS WITH RANDOM YIELDS AND LEAD TIMES

Candace Arai Yano
Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, MI 48109-2117

Technical Report 92-54

October 1992

INPUT CONTROL FOR SYSTEMS WITH RANDOM YIELDS AND LEAD TIMES

Candace Arai Yano
Department of Industrial & Operations Engineering
University of Michigan
Ann Arbor, MI 48109-2117*

September 1992

*Visiting at IE-EM Department, Stanford University, Stanford, CA 94305-4024 from August 1992 until July 1993. A portion of this work was done while the author was visiting the IEOR Department at the University of California, Berkeley.

INPUT CONTROL FOR SYSTEMS WITH RANDOM YIELDS AND LEAD TIMES

Abstract

We analyze a class of input control policies for dealing with the effects of random production yields and the consequent effects of the random yields on the lead times. The policy has the following form: an input quantity is selected at the beginning of each time period to satisfy, with a specified probability (service level), the (cumulative) demand between now and one planned lead time hence, taking into account current only hand inventory and any in-process production batches.

We obtain analytic expressions for the first three moments of the input quantities and inventory levels that arise from such a policy under the assumption that the lead time is one period or less. We imbed these results within a queueing model, which enables us to capture the effects of the congestion caused by the random yields. To analyze situations when the planned lead time is greater than one period, we developed a simulation model because exact results cannot be obtained. The results suggest that the variance of the batch sizes actually decreases as the planned lead time increases, because of negative autocorrelation among the sizes of the batches in process. These results provide the basis for a simple iterative procedure to determine the smallest planned lead time that is consistent with a given service level.

INPUT CONTROL FOR SYSTEMS WITH RANDOM YIELDS AND LEAD TIMES

1. INTRODUCTION

The process of manufacturing (electronic) wafers is characterized by both random yields and long and uncertain flow times. The long flow times are often a consequence of the strategic decision to maintain high work-in-process inventory levels to maintain high levels of throughput. This congestion in the system is largely responsible for the uncertainty in the flow times.

This combination of random yields and lead times creates the need for control policies that can *compensate*, to some extent, for both types of uncertainty simultaneously, and thereby maintain acceptable levels of customer service. At the same time, it would be desirable to develop polices that will *limit* the adverse effects of random yields on the mean and variance of the lead time.

In this paper, we analyze a class of two-parameter control policies that resemble those we have observed in practice. The policies have the following form: Decisions are made at the beginning of each time period, where the duration of the time period is determined exogeneously. The input quantity is released a fixed period of time (i.e., a planned lead time) prior to the desired completion time of the batch, where the duration of this time offset is a function of the distribution of the actual flow time (also referred to as the lead time). The input quantity is a constant multiplied by net demand (or the net production target, if a production smoothing mechanism is employed) in the period one planned lead time hence, where the constant scaling factor is a function of the yield distribution.

We are aware of one instance where the company uses a planned lead time equal to the mean lead time and a constant scaling factor equal to the reciprocal of the mean yield rate (average fraction good), and another where the company uses a planned lead

time equal to the 95th percentile of the lead time distribution, and a constant scaling factor equal to the reciprocal of the 5th percentile of the yield rate distribution (thereby ensuring that the production target is satisfied 95% of the time).

One important feature of this class of policies is its ease of implementation. Relatively little data are required to choose parameters that are consistent with the organization's objectives, and the scheduler only needs to know two parameters to execute the policy. Another important feature is that this class of policies is that various performance characteristics can be obtained, making it possible to compute good parameter values for a variety of different objective functions. Of course, a scheduler may choose or need to deviate from these policies for a variety of different reasons, such as capacity constraints, rush orders, etc. In this paper, our concern is not to develop real-time control policies, but to develop an understanding of the characteristics of policies of this type to provide a basis for setting better guidelines for long-term strategies.

Despite the many attractive features of the policies described above, the analytical models required to capture various performance measures are not simple. Consequently, in this paper, we limit our investigation to single-product systems and treat the production system as a "black box"--in essence, a single stage system. Admittedly, this approximation of the true system is crude, but it allows us to examine in detail some issues and interactions that are easily obscured when more complicated models of the manufacturing process are used. Moreover, the schedulers at the two companies mentioned above are, in fact, treating the production system as a black box in much the same way as we are here. Thus, we believe that even these gross approximations will not limit the transferability of our results into practice.

We are not aware of any literature that treats both the "lot sizing" and the resultant queueing aspects (in the sense of waiting time *distributions*) of our model, where uncertain yields are the source of uncertainty. We are aware of only one paper that has several similarities: a paper by Courcoubetis et al. (1989), who consider a multi-product,

multi-machine scheduling problem where all machines can produce all products, but with differing probabilities of success. In their model, orders arrive according to a stochastic process, whereas in our model, production is performed in anticipation of demand. They assume that a machine, when assigned to produce a product, will continue to produce at least until any backlogged demand is satisfied. They develop necessary and sufficient conditions for existence of a scheduling policy that produces bounded inventory (or backorder) levels, and develop such a scheduling policy. They do not deal directly with the distributional aspects of the lead time.

Most other papers are only tangentially related to our research. There are a variety of control-theory-based real-time control approaches for flow control in manufacturing systems (e.g., Gershwin et al. 1985; Gong and Matsuo 1989, 1990). Most of these approaches have the objective of tracking the demand trajectory (minimizing weighted shortages and overages or an analogous objective), and decisions are made using detailed information about the locations and quantities of work-in-process inventory. With very few exceptions, lead times are not considered explicitly, and when they are, they are treated only in an average sense. One paper that considers lead times is by Graves (1989), who analyzes policies in which the production policy is to process, in each period, a fixed percentage of the work in the input queue. The percentage is selected to be the reciprocal of the planned lead time. Karmarkar (1991) provides a thorough overview of lead time management issues in manufacturing settings.

Research has also been done on the control of queues (e.g., Federgruen and Groenevelt 1988a,b; Shanthikumar and Yao 1992 and references therein). Our problem differs from standard queueing (or queueing network) models in that there are no arrivals in our problem: we must determine the release quantities to satisfy some future demand. Altenatively, our system can be viewed as a queue in which we have some information about future arrivals, and where those future arrivals expect to be served instantaneously from stock on hand. These systems have been referred to as "make-to-stock" queues.

Such systems have been studied for decades (see references in Wein 1992a,b, and in Chapter 4 of Buzacott and Shanthikumar 1992), but the number of papers in the literature is still relatively limited, and only a few of these papers capture yield randomness. Ou and Wein (1991) develop an approach for a multi-class feedback (reentrant) queue in which releases are controlled (in an average sense), but matching a steady-state target throughput (minimizing inventory and backorders) is the primary consideration and due dates are not considered explicitly.

Gong and Matsuo (1990a,b) analyze linear control rules for determining release quantities. They formulate a control problem with the objective of minimizing the weighted variance of work-in-process and production levels, and present a method for obtaining weights that minimize the expected total work-in-process inventory while ensuring that production capacity constraints are satisfied with a specified probability. In the second paper, this approach is extended to allow for re-entrant routings, more general yield distributions, and rework. Although production constraints are considered to some extent in their model, they do not explicitly analyze the variability of waiting times.

There is also a growing literature on single-stage, multi-period, random yield lot sizing models (see Yano and Lee 1989 for a review). With few exceptions, these papers do not consider capacity constraints, and those that do consider capacity constraints do not consider queueing effects.

In the next section, we detail our problem assumptions and describe our model formally. We then provide a complete development of a model for the case where the planned lead time is assumed to be one period. We use this as the basis for analyzing systems in which the planned lead time may be greater than one period. In section 5, we discuss possible extensions of our model, and conclusions appear in section 6.

2. ASSUMPTIONS AND MODEL DEVELOPMENT

We consider a situation with a single "machine" or system that produces a single product. The product experiences random yield losses in the manufacturing process, but the extent of yield loss is not known until the product is inspected at completion (or at some other specified time in the process). We assume that the distribution of the fraction good from each periodic release is independent of the release quantity, and is stationary over time. For simplicity, we assume that there is a periodic, constant demand for the product. (Extension to stationary demand processes is conceptually straightforward but algebraically messy.) Many semiconductor manufacturing facilities produce to production targets (at least partially based on demand forecasts, but with some production smoothing) rather than actual orders, and these targets tend to be fairly stable. Thus, the assumption of constant demand is not unrealistic.

At the beginning of each time period, a quantity is released to the system which is intended to satisfy the demand one planned lead time hence. We assume that the production batches are processed in first-come-first-served (FCFS) sequence; consequently, we assume that orders do not cross. The FCFS service discipline is easily enforced in a single-server system. We also assume that demands are filled in FCFS sequence. In each period, we require the release quantity to be sufficiently large so that there is a high probability of satisfying all of the demand, taking into account the *distribution* of the on-hand inventory (and backorders) upon completion of production of the batch.

We next introduce the notation that will be used throughout the paper. The symbol ~ denotes 'with distribution function.'

D = demand per period.

P = yield rate random variable ~ F.

p_n = yield rate in period n (fraction of good product).

I_n = end-of-period-inventory in period n.

a = yield adjustment factor.

α = service level (probability of satisfying demand assuming that the batch is completed within the planned lead time).

h = inventory holding cost per unit per period

π = shortage cost per unit per period.

Q_n = input quantity of period n.

W = flow time (random variable).

W_q = waiting time in queue (random variable).

ρ = utilization level.

c_s^2 = squared coefficient of variation of the service time.

σ_s^2 = variance of the service time.

3. MODEL WITH LEAD TIME LESS THAN ONE PERIOD

In this section, we assume that the production lead time is less than one period with probability 1. Later in the section we derive conditions in which this assumption is valid with high probability. We consider a production policy of the form

$$Q_n = a\,(D - I_n)$$

which in turn ensures that

$$P[I_{n-1} + p_n Q_n \geq D] = \alpha.$$

where $\alpha = 1 - F(a^{-1})$. This is essentially a service level policy which ensures that demand is satisfied with probability α *in each period.* For the moment, we assume that $D - I_{n-1}$ is always non-negative. Later in the section we derive conditions on P that ensure this condition is true with a high probability under the stated production policy.

If we start with $I_0 = 0$, then it is easy to show that

$I_n = (1 - ap_n)(I_{n-1} - D)$.

Using the fact that p_n and I_{n-1} are independent, it can be shown (see Appendix) that

$$\lim_{n\to\infty} E(I_n) = \frac{D(aE(P) - 1)}{aE(P)}$$

if $|1 - aE(P)| < 1$, which is equivalent to $\frac{1}{E(P)} < a < \frac{2}{E(P)}$, where it is assumed without loss of generality that $E(P) > 0$. This condition simply means that the yield adjustment factor is between one and two times that required to compensate for the average yield loss. In most cases, this is a very mild condition. Economic factors (e.g., shortage costs greater than inventory holding costs) ensure that the lower limit is not binding, and capacity constraints are often much more restrictive than the constraint imposed by the upper limit. Throughout the remainder of this paper, we will simply assume that this condition holds.

With the above results, it can also be shown that under certain relatively mild conditions (see Appendix), $E(I_n^2)$ converges as $n \to \infty$, and that

$$\lim_{n\to\infty} E(I_n^2) = D^2 \frac{[2-aE(P)][1-2aE(P)+a^2E(P^2)]}{a^2E(P)[2E(P)-aE(P^2)]}$$

and that

$$\lim_{n\to\infty} \mathrm{Var}(I_n) = \frac{D^2\{E(P^2)-[E(P)]^2\}}{a\,[E(P)]^2\,[2E(P) - aE(P^2)]}$$

In addition, under certain conditions on the first three moments of P (see Appendix A), we also have

$$\lim_{n\to\infty} E(I_n^3) =$$

$$D^3 \frac{[a^3E(P^3)-3a^2E(P^2)+3aE(P)-1][a^3E(P)E(P^2)-2a^2[E(P)]^2-3a^2E(P^2)+9aE(P)-6]}{a^2E(P)[2E(P)-aE(P^2)][-3aE(P)+3a^2E(P^2)-a^3E(P^3)]}$$

The third moment may be useful in characterizing the distribution of I, which might be useful if one would like to optimize a, as we discuss later in the paper.

Using the mean and the variance of I, we can easily derive conditions in which D-I is non-negative with a high probability. We assume that these conditions are satisfied if

$$E(I) + 2\sigma_I \leq D,$$

where σ_I is the standard deviation of I. Now both E(I) and σ_I are increasing with a; thus, $E(I) + 2\sigma_I$ is strictly increasing with a. (Var(I) is decreasing with a if $a < E(P)/E(P^2)$ and increasing with a otherwise. Note however, that since $Var(P) > 0$, we must have $E(P^2) > [E(P)]^2$, which implies that $E(P)/E(P^2) < E(P)^{-1}$. Thus, if $a < E(P)/E(P^2)$, then $a < E(P)^{-1}$, which violates a basic assumption. Consequently, we have Var(I) increasing with a over the relevant range.) A reasonable upper limit on $E(I) + 2\sigma_I$ is achieved when $a = (\mu_P - 2\sigma_P)^{-1}$, i.e., when a is set equal to the reciprocal of the yield fraction that is two standard deviations below the mean yield rate. Letting $b = \sigma_P / \mu_P$, we have (after some algebra) that

$$E(I) + 2\sigma_I = 2Db(1-2b)[(1-2b)^{-1} + (1-5b)^{-0.5}],$$

which is less than D for values of b less than roughly 0.175.

More generally, defining $a = (\mu_P - c\sigma_P)^{-1}$ for some non-negative value of c, we have $a = [\mu_P (1-bc)]^{-1}$, and it is easily shown that

$$E(I) = bcD$$

and $$Var(I) = \frac{b^2(1-bc)^2D^2}{1-2bc-b^2}.$$

With these results, it is straightfoward to show that $E(I) + 2\sigma_I \leq D$ if

$$c \leq (1-5b^2)/2b. \qquad (1)$$

This does not impose any practical constraints for $b \leq 0.2$, and imposes only mild constraints on c for values of b up to 0.3. Thus, if the coefficient of variation of the yield rate distribution is not too large, and if we do not insist upon a very high service level (large value of c) when it is, we will have $I \leq D$ a great majority of the time.

Flow Time Analysis

We now turn to an analysis of the distribution of the flow time. Since we have the first two moments of I, we can readily obtain the first two moments of Q. Observe, however, that under our release policy, the Q's are not independent but are negatively autocorrelated. Because of the limited results on queues with autocorrelated arrivals, however, incorporating the autocorrelation would be extremely difficult. We therefore analyze the system as if the Q's were independent, realizing that both the mean and the variance of the flow time is likely to be less in the "real" system. Later in this section, we report experimental results on the bias caused by this approximation.

If the processing time per unit is deterministic, say τ per unit, then the processing time of the batch has the same distribution as that of τQ. On the other hand, if the processing time per unit is stochastic, we could simply convolute this processing time distribution with the distribution of Q to obtain the distribution of the service time per batch. For ease of exposition and because our primary interest is in analyzing the effects of the yield rate distribution on flow times, we will treat the case where the processing time per unit is deterministic. However, since the batch sizes are random, the processing time per batch is also random.

Since we are considering a periodic release policy, our system can be modeled as a D/G/1 queue. If we let the interarrival time be one time unit ("period") then the service time can be expressed as the fraction of the period's capacity required to process the batch. With these definitions, exisiting results on D/G/1 queues can be applied directly. Unfortunately, few results exist. Servi (1986) derives the z-transform of the waiting time, from which we can obtain the various moments, but not the complete distribution of the waiting time. For planning purposes, the mean and variance might be sufficient, as we will show later. We should point out that while a variety of other results are available to estimate or bound the mean and variance of the waiting time (e.g., Kleinrock 1976, Tijms

1986), most require the complete distribution of the service time, but we have only the moments. Thus, we limit ourselves to results that require only moments of the service time distribution.

Recall that one of the reasons for this queueing analysis is to ascertain conditions in which the assumption of a one-period lead time is true with high probability. (Note that we can guarantee that the assumption is true with probability 1 only if the maximum value of Q is bounded, but this cannot be assured under our control policy. In principle, an infinite value of Q may occur, albeit with infinitesimal probability.)

To assess conditions in which the lead time (flow time) is less than one period with high probability, we can use the results of Servi (1986) mentioned above. To use these results, we first need to obtain the first three moments of the service time distribution. For our problem, we have (after some algebra)

$$E(Q) = D/E(P) \tag{2}$$

and

$$E(Q^2) = \frac{D^2[2-aE(P)]}{E(P)\,[2E(P)-aE(P^2)]} \tag{3}$$

Thus,

$$Var(Q) = \frac{aD^2\{E(P^2)-[E(P)]^2\}}{[E(P)]^2\,[2E(P) - aE(P^2)]}.$$

It is interesting to note that Var(Q) is increasing with a. Thus, although increasing a will generally improve customer service with respect to quantity, the price to be paid is a higher variance in the production quantity, which in turn increases the variance of the flow time.

We also have

$$E(Q^3) = a^3D^3\{1 - \frac{3[aE(P)-1]}{aE(P)} + \frac{3[2-aE(P)][1-2aE(P)+a^2E(P)]}{a^2E(P)[2E(P)-aE(P^2)]}$$

$$- \frac{[a^3E(P^3)-3a^2E(P^2)+3aE(P)-1][a^3E(P)E(P^2)-2a^2[E(P)]^2-3a^2E(P^2)+9aE(P)-6]}{a^2E(P)[2E(P)-aE(P^2)][-3aE(P)+3a^2E(P^2)-a^3E(P^3)]}\}.$$

Specializing the results of Servi to our case (with an interarrival time of one period), we have

$$E(W) = \frac{G''(1)}{2[1-G'(1)]}$$

and $$\text{Var}(W) = \frac{G'''(1)}{3[1-G'(1)]} + \frac{G''(1)}{2[1-G'(1)]} + \{\frac{G''(1)}{2[1-G'(1)]}\}^2$$

where $G(\cdot)$ is the z-transform of the service time distribution, and where

$G'(1) = E(\tau Q)$,

$G''(1) = E(\tau^2 Q^2) - E(\tau Q)$,

$G'''(1) = E(\tau^3 Q^3) - 3E(\tau^2 Q^2) + 2E(\tau Q)$.

If both E(W) and Var(W) are sufficiently small, then we may be able to conclude that the probability of the flow time being greater than 1 is small. We should note, however, that the derivations of E(W) and Var (W), as well as the moments of I and Q, are based on the assumption that the lead time is less than one period. If this assumption is not satisfied, the expressions given above may not provide accurate estimates if delayed feedback about yield outcomes (due to the lead time) increases the variance of Q. We would expect the estimates to be useful for practical purposes if there is only a small probability that the flow time is greater than one. Another somewhat more robust method to address the same issue is discussed below.

An approximation of $P\{W_q > 0\}$ has been developed by Kramer and Lagenbach-Belz (1976) for GI/G/1 queues. Specializing the approximation to our case (with deterministic interarrival times), we have

$$P\{W_q > 0\} = \frac{(\rho^2 + \rho^4)\, c_s^2}{(1-\rho) + (\rho + \rho^2)\, c_s^2}.$$

where c_s^2 is the squared coefficient of variation of the service time. We have tabulated this probability for several combinations of ρ and c_s^2 in Table 1. These results are helpful in specifying when the lead time is less than 1 with high probability. In order to set planned lead times, however, it would be useful to have more information about the tail of the waiting time distribution. It should be clear from Table 1 that in light traffic, say $\rho \leq 0.6$ and with moderate squared coefficients of variation of the service time, say $c_s^2 \leq$

0.1, a very small percentage of the batches will have a lead time of more than one period. In other instances, however, the lead times may be longer. In the remainder of this section, we show that under reasonable conditions, the probability of a lead time exceeding 3 or 4 periods is quite small when the moments of Q are as speecified in (2) and (3). In the next section, we make use of this fact to analyze for the case of lead times greater than one period.

Table 1

As noted above, the probability of a lead time greater than 1 is fairly small except in heavy traffic, or in moderate traffic with a high squared coefficient of variation of the service time. Consequently, we decided to use a heavy traffic approximation (Kleinrock 1976, p. 31) to estimate the tail probabilities for the waiting time. Again, specializing the formula to our specific case, we have:

$$W_q(y) \approx 1 - \exp(-\frac{2(1-\rho)y}{\sigma_S^2}).$$

In Figure 1, we illustrate regions in which $P\{W_q > k\}$ is less than 0.01, 0.05, and 0.10 for $k = 1,...,4$. Note that the mean service time is equal to ρ here so $\sigma_S^2 = 0.1$ ($\sigma_S = 0.32$) is quite large in comparison. If $\sigma_S^2 \leq 0.1$ and $\rho \leq 0.9$, we can easily ensure that $P\{W_q > 2\} \leq 0.05$. Similarly, if $\sigma_S^2 \leq 0.1$ and $\rho \leq 0.95$, we can ensure that $P\{W_q > 3\} \leq 0.05$. Thus, for practical purposes, it appears that we can use a maximum planned lead time of 3 or possibly 4 periods, thereby simplifying the analysis in the next section.

4. ANALYSIS OF SYSTEMS WITH LEAD TIME GREATER THAN ONE PERIOD

In this section we consider the same type of input control policy as in the previous section, but allow the planned lead time to be greater than one period. We assume that the batch will be completed within the planned lead time with a high probability, and that the yield is observed after the planned lead time has elapsed. Of course, in reality, the batch may be completed and consequently, the yield may be observable before the

planned lead time has elapsed. It is easy to show that the probability is quite small that a batch will be completed an entire period earlier than planned. For example, if $\sigma_s^2 = 0.1$ and $\rho = 0.9$, we might choose a planned lead time of two periods (because $P\{W_q > 2\} \leq 0.05$), and if $\sigma_s^2 = 0.1$ and $\rho = 0.95$, we might choose a planned lead time of three periods (because $P\{W_q > 3\} \leq 0.05$). In either case, only about 13.5% of the batches will be completed a full period "early." Moreover, we need to remember that the impact of these "early" completions is partially offset by the late completion of 5% of the batches. Consequently, for practical purpose, it seems reasonable to assume that the yield will be observed after the planned lead time has elapsed. It is also useful to point out that since a batch is released only once per period, there is no advantage in using fractional (rather than integral) planned lead times from the standpoint of providing more rapid updating of the input quantities based on observed yields. Faster feedback may help to improve quality, however.

With these additional assumptions, we have the following input policy: choose Q_n such that

$$P\{P_{t-L+1}Q_{t-L+1} + P_{t-L+2}Q_{t-L+2} + \cdots + P_tQ_t + I_{t-1} \geq LD\} \geq \alpha$$

where we redefine P_t as the yield rate for the batch started in period t. Because it is not possible to write a closed-form expression for Q_n, it is also not possible to derive analytic expressions for the mean and variance of the flow times. Fortunately, the approximation of Kramer and Lagenbach-Belz and that of Kleinrock do not require very accurate information about c_s^2 or σ_s^2 if we only desire to set planned lead times appropriately. Indeed, if we could make some statements about the distribution of Q for a given planned lead time, it would be relatively easy analyze the problem in an iterative fashion. One possible iterative scheme follows.

We would start with an initial planned lead time using an approximation for the tail of the waiting time distribution, where the service time variance is based on the (incorrect) assumption that the lead time is less than one period. Then, we could obtain

more accurate estimates of the moments of Q for the initial planned lead time. If the moments are similar to those previously computed, then we are done. Otherwise, we would use the new moments to select a new planned lead time, and repeat the process until the planned lead time decision converges. Since the planned lead time decision is not sensitive to the exact values of the moments of Q, such a procedure should perform well.

The problem at this point is to develop a method to estimate the moments of Q for a given planned lead time. One question that arises in this context is whether the moments are similar to those obtained under the assumption of a one-period lead time. Lengthening the lead time might be expected to increase the variance of Q due to the uncertainty of the yields of the in-process batches. On the other hand, it is important to remember that the sizes of the batches "in-process" are negatively autocorrelated, and this may partially or fully compensate for the increased level of uncertainty.

To investigate this issue, we developed a simulation model to generate a series of Qs, from which we could estimate the first two moments. We used a variety of different lead times, service levels, and yield rate distributions, as shown in Table 2. For service levels of 80% and 90%, we compare the first two moments of Q estimated from the simulation with the analytic expressions given in Section 3, under the assumption that $\alpha = F(a^{-1})$.

Table 2

It is evident from the results that the variance of the batch size (and therefore the variance of the service time) generally *decreases* with L due to the negative autocorrelation among the Q_ns for lags up to L-1. (Indeed, any discrepancies from this pattern appear to be due to statistical fluctuations. We found that long simulation runs were needed even to ensure that the sample mean for Q was within 1% of the true mean, and even small discrepancies in the sample mean can cause noticeable discrepancies in the higher moments. We initially used several dozen observations but observed large

discrepancies in the sample mean. Consequently, we decided to report results based on 200 observations. Longer simulation runs would have provided slightly smoother declines in the service time variances, but would not have affected the overall conclusions.)

These results suggest that using the service time variance for the case of L=1 will give conservative (longer) estimates of the 'best' planned lead time. Moreover, since the variance of the service time actually declines (rather than increases) with L, the iterative scheme described earlier should perform well. Starting with the variance based on L = 1, one will initially overestimate the planned lead time required to ensure a specified level of on-time service. This, in turn, leads to an decrease in the estimate of the variance of the service time and a consequent decrease in the estimate of the planned lead time. In the next iteration, the two estimates will increase from the values in the previous iteration, and in the subsequent iteration, they will decrease. In principle, oscillatory or divergent behavior may occur, but since only a small number of lead times are under consideration, we would not expect it to present practical difficulties.

We implemented the iterative scheme for all of the distributions and service levels in Table 2, at values of $\rho = 0.8$, 0.9, and 0.95. All of these cases satisfy the conditions on the coefficient of variation of the yield distribution and the service level given by (1). In all cases, the solution could be obtained in at most three iterations, with a lead time of one period in most cases, a lead time of two periods in a few cases with high service levels and high levels of utilization, and a lead time of three periods in one case with a high service level, high level of utilization, and the highest variance of the yield rate in our data set. (In two cases, the solution procedure oscillated between two consecutive lead time values. In such cases, a manager would need to decide between a shorter planned lead time with a lower on-time percentage or a longer planned lead time with a higher on-time percentage.)

It appears that the proposed control policy gives very acceptable planned lead times, even at relatively high levels of utilization.

Although in most cases, the iterative procedure terminated quickly, when the trial lead time was greater than one period, simulation had to be used to estimate the service time variances. Further research is needed to develop fast, approximate estimates of the service time variance for lead times greater than one period so that simulation can be avoided altogether.

Our model was not designed to represent cases where the condition in (1) is not satisfied. Nevertheless, we were interested in determining whether such systems were actually unstable, or whether they were simply not represented accurately by our models. For this purpose, we ran a few simulations with yield distributions and service levels that do not satisfy the condition that the on-hand inventory is less than D with a high probability. We observed that the variances of the batch sizes obtained from our model for the case of L = 1 overestimated the true variances by 20% up to tenfold in our examples. Thus, it would not be wise to use our model to estimate the variances in these circumstances. The observed coefficients of variation are in the 0.5 to 0.6 range, and as in the other cases, decline as the lead time increases. While such variability would be undesirable in practice, when necessary, it might be possible to accommodate it at moderate levels of utilization and with appropriately chosen lead times.

In the simulation results cited above, we computed exact values for the Qs, but found the process to be very CPU intensive. In practice, the computational burden of finding the exact values of Q_t simply to estimate the variance may not be justified since an accurate estimate of the variance is not needed. For example, a even a 10% error is unlikely to change our choice of the planned lead time. As an alternative, we could generate a series of approximate values of Q_t as follows. Observe that

$$\sum_{i=1}^{L} P_{t-L+i}\, Q_{t-L+i}$$

has mean

$$\mu_P \sum_{i=1}^{L} Q_{t-L+i}$$

and variance

$$\sigma_P^2 \sum_{i=1}^{L} Q_{t-L+i}^2$$

for any arbitrary vector of Qs. As an approximation, we could choose Q_t such that

$$\frac{I_{t-1} + \mu_P \sum_{i=1}^{L} Q_{t-L+i} - LD}{\sigma_P \left[\sum_{i=1}^{L} Q_{t-L+i}^2 \right]^{0.5}} = t_{L,\alpha}$$

where $t_{L,\alpha}$ is the Student's t-statistic corresponding to the α percentile and L degrees of freedom. Our reason for considering a t-statistic rather than a normal approximation is that we expect L to be relatively small (four or less).

5. EXTENSIONS

Optimizing the Value of *a* for L=1

Thus far, we have assumed that the value of α is given, and consequently also the value of *a*. In some settings, α may not be specified *a priori*, providing an opportunity to choose an economically optimal value of *a*. We could obtain a good value of *a* by solving the following optimization problem:

minimize $h\, E(I^+ \mid a) + \pi\, E(I^- \mid a)$

where h is the inventory holding cost per unit per period, π is the per unit shortage cost per period, $I^+ = \max(I, 0)$ and $I^- = -\min(I, 0)$. Note that if shortages are likely to be small (e.g., when there is a high π/h ratio), we might choose to approximate the first term by $hE(I \mid a)$, which simplifies the problem considerably. In this case, it would be sufficient to obtain a good fit for the left hand tail of the density (where inventory is negative).

The objective function may not be convex for all yield distributions. Indeed, since Var(I) is increasing with a in the relevant range, there may be multiple local minima. That is, increasing a may not decrease expected shortages even though the average inventory level is increased. However, like most other inventory problems, we would expect the objective function to be relatively well-behaved. Consequently, a simple search procedure should produce very good solutions. We will illustrate this point by an example.

In the following example, the yield distribution is assumed to be normal with a mean of 0.8 and a standard deviation of 0.05. We approximate the distribution of I by using the first two moments from our analytic model and assuming the distribution is normal. We assume D = 100, h= 1 and π = 10. We approximate the first term in the objective function by hE(I | a), which should be reasonably accurate because shortages are quite costly, and thus the expected shortage quantity should be small in a reasonable solution. The expected costs are tabulated below for a range of values of a:

α	a	Total expected cost
0.8	1.319	12.20
0.85	1.336	11.38
0.875	1.346	11.19
0.9	1.359	11.01
0.925	1.373	11.17
0.95	1.393	11.63
0.98	1.434	13.33

In this example, the optimal service level is in the neighborhood of 90% and the objective function seems to be well-behaved.

Optimizing a for $L > 1$

Earlier in the paper we described an iterative procedure that can be used to determine the planned lead time for any given value of a. The challenge in optimizing a for $L > 1$ lies in the fact that we do not know L *a priori*. Increasing a ensures a higher probability of satisfying the demand and leads to an increase in safety stock. It may also lead to a reduction in shortages. Because of the queueing effects in our model, increasing a also increases the variance of the lead time, and therefore may also increase the planned lead time. Since work-in-process inventory is proportional to the planned lead time, its cost should be considered in selecting the value of a. The optimization problem becomes:

$$\underset{a}{\text{minimize}}\ h_f E(I^+ \mid a) + h_w D\, L^*(a) + \pi E(I^- \mid a)$$

where h_f = inventory holding cost per unit per period for finished goods inventory

h_w = inventory holding cost per unit per period for work-in-process inventory

$L^*(a)$ = shortest planned lead time that satisfies the specified on-time probability for a given a

Solving this problem is not straightforward because $L^*(a)$ is a step function. With reasonable approximations such as those described earlier in this section, a judicious search over a range of values of a might produce reasonable solutions. It is useful to point out that since the second term is non-decreasing in a, the optimal value of a in this problem is strictly smaller than the value that optimizes the corresponding problem for L = 1. The problem that remains is estimating the variance of I for a variety of service levels and lead times. Further research is needed to address this issue.

6. SUMMARY AND CONCLUSIONS

We have analyzed a class of input control policies that resemble those we have observed in practice for dealing with the effects of random production yields and the consequent effects of the random yields on the lead times. The policy has the following

form: an input quantity is selected at the beginning of each time period to satisfy, with a specified probability (service level), the (cumulative) demand between now and one planned lead time hence, taking into account current only hand inventory and any in-process production batches. The choice of the service level affects the variance of the input quantity, which in turn affects the variance of the flow time, which then affects the choice of the planned lead time. Our main objective in the paper was to develop an understanding of this complicated relationship.

We have obtained analytic expressions for the first three moments of the input quantities and inventory levels that arise from such a policy under the assumption that the lead time is one period or less. We imbed these results within a queueing model, which enables us to capture the effects of the congestion caused by the random yields. This then allows us to characterize when this assumption regarding the lead time is satisfied with a high probability.

When the planned lead times are greater than one period, it is not possible to write closed-form expressions for the various moments of relevance. To understand how the variance of the input quantity changes with the duration of the planned lead time, we developed a simulation model. The results suggest that the variance of the batch sizes actually decreases as the planned lead time increases, because of negative autocorrelation among the sizes of the batches in process. These results provide the basis for a simple iterative procedure to determine the smallest planned lead time that is consistent with a given service level.

We also briefly discuss the problem of optimizing the service level, taking into account finished goods, work-in-process inventory, and shortage costs. Further research is needed to develop an efficient way to solve this problem when the planned lead time is greater than one period. Research is also needed to generalize models of this type to systems with multiple production stages. One other interesting generalization would be to relax the (hard) service level constraints by allowing for partial response to low yield

outcomes and by allowing the system to 'produce ahead' when the input quantity would otherwise be low. Such policies could reduce the variance of the input quantity, potentially leading to shorter planned lead times in high service level, high utilization instances.

Acknowledgement

This research was supported in part by National Science Foundation Grant HRD-91-03229 to the University of California (Berkeley), and National Science Foundation Grant EID-90-23674 to the University of Michigan. The author is grateful for the assistance of Mr. Jeenyoung Kang who coded the simulation model and obtained the simulation results.

References

Buzacott, J.A. and J.G. Shanthikumar, Stochastic Models of Manufacturing Systems, Prentice-Hall, Inc., Englewood Cliffs, NJ. 1992.

Federgruen, A. and H. Groenevelt, "M/G/c Queueing Systems with Multiple Customer Classes: Characterization and Control of Achievable Performance under Nonpreemptive Priority Rules," *Management Science*, Vol. 34, No. 9, pp. 1121-1138, 1988a.

Federgruen, A. and H. Groenevelt, "Characterization and Optimization of Achievable Performance in Queueing Systems," *Operations Research*, Vol. 36, No. 5, pp. 733-741, 1988b.

Gershwin, S.B., R. Akella, and Y. Choong, "Short Term Production Scheduling of an Automated Manufacturing Facility," *IBM Journal of Manufacturing and Development*, Vol. 29, No. 4, pp. 392-400, 1985.

Gong, L. and H. Matsuo, "Stabilizing Work-in-Process and Smoothing Production in a Production System with Random Yield, " Working Paper, Graduate School of Business, University of Texas, Austin, TX, 1990a.

Gong, L. and H. Matsuo, "A Control Policy for a Manufacturing System with Random Yield and Rework, " Working Paper, Graduate School of Business, University of Texas, Austin, TX, 1990b.

Graves, S.C., "Safety Stocks in Manufacturing Systems," *Journal of Manufacturing and Operations Management*, Vol. 1, No. 1, pp. 67-101, 1988.

Karmarkar, U., "Manufacturing Lead Times, Order Release and Capacity Loading," in Logistics of Production and Inventory, S. Graves, A.H.G. Rinooy Kan, and P. Zipkin (eds.), Handbook in Operations Research and Management Science, Vol. 4, North Holland, 1991.

Kleinrock, L., Queueing Systems (Vol. II), John Wiley & Sons, Inc., 1976.

Kramer, W. and M. Langenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," in *Proceedings of the 8th International Teletraffic Conference*, Melbourne, p. 235, 1976. Also described in Kleinrock, *op. cit.*

Lee, H. and C. A. Yano, "Production Control for Multi-Stage Systems with Variable Yield Losses," *Operations Research*, Vol. 36, No. 2, 1988.

Ou, J. and Wein, L.M., "Dynamic Scheduling of a Production/Inventory System with By-Products and Random Yield," Working Paper OR-250-9, Operations Research Center, Massachusetts Institute of Technology, 1991.

Servi, L. D., "D/G/1 Queues with Vacations," *Operations Research*, Vol. 34, No. 4, 1986.

Shanthikumar, J.G. and D.D. Yao, "Multi-Class Queueing Systems: Polymatroidal Structure and Optimal Scheduling Control," *Operations Research*, Vol. 40, Supplement 2, pp. S293-S299, 1992.

Tijms, H.C., Stochastic Modeling and Analysis, John Wiley & Sons, 1986.

Wein, L.M., "Dynamic Scheduling of a Multiclass Make-to-Stock Queue," *Operations Research*, Vol. 40, No. 2, pp. 724-735, 1992a.

Wein, L.M., "Scheduling Networks of Queues: Heavy Traffic Analysis of a Multistation Network with Controllable Inputs," *Operations Research*, Vol. 40, Supplement 2, pp. S312-S334, 1992b.

Yano, C. A. and H. Lee, "Lot-Sizing with Random Yields: A Review'" Technical Report 89-16, University of Michigan, Ann Arbor, 1989 (to appear in *Operations Research*).

Table 1

$P(W_q > 0)$ from the Approximation of Kramer and Lagenbach-Belz

		c_s^2				
		0.01	0.05	0.10	0.20	0.30
	0.2	0.0002	0.0026	0.0051	0.0098	0.0143
	0.4	0.0031	0.0148	0.0283	0.0521	0.0725
ρ	0.6	0.0120	0.0546	0.0987	0.1654	0.2134
	0.8	0.0490	0.1929	0.3051	0.4302	0.4982
	0.9	0.1252	0.3952	0.5410	0.6634	0.7175

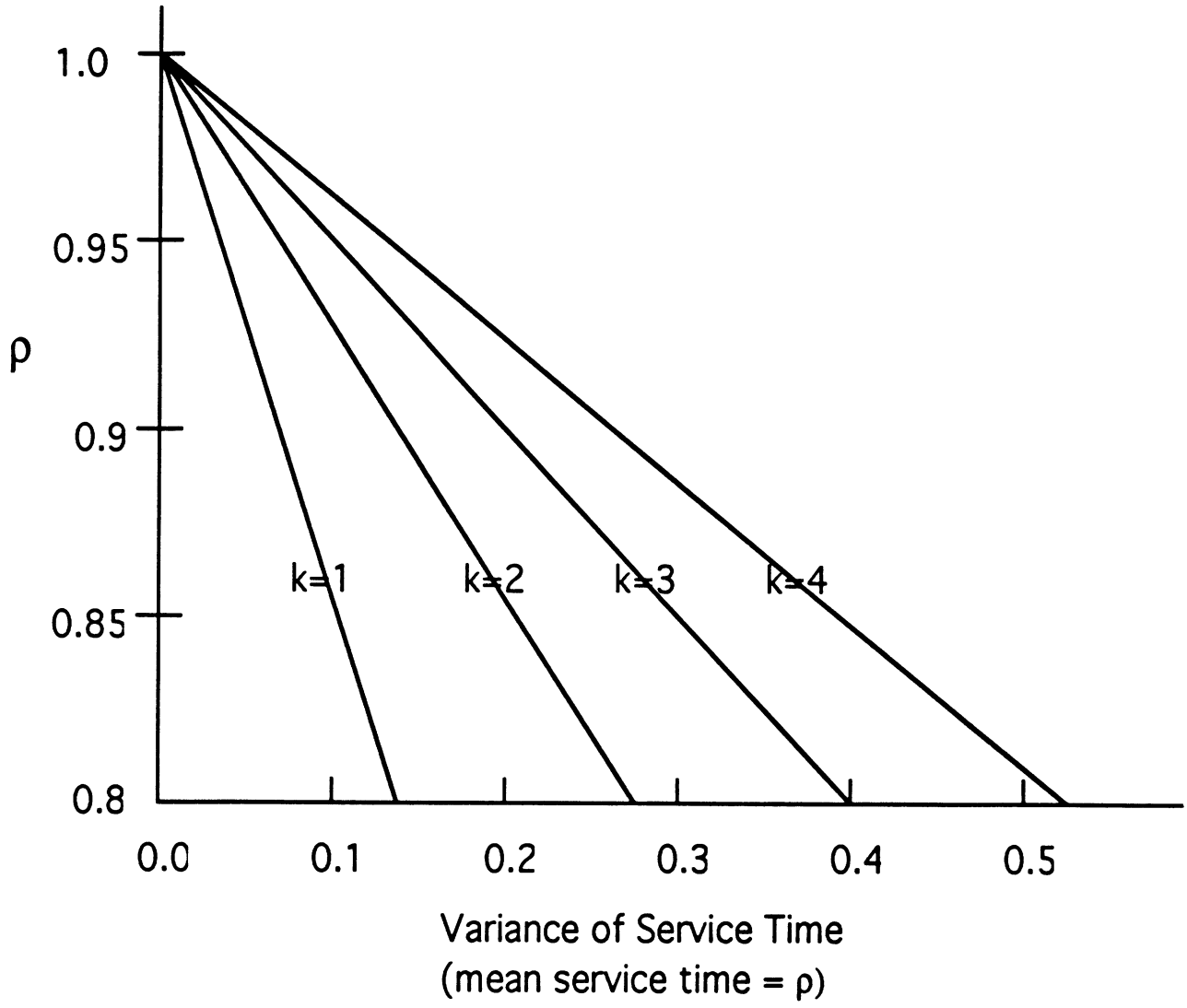

Figure 1: Boundaries of regions in which $P\{W_q > k\} < 0.05$

Table 2

Comparison of Var(Q) for L=1 (exact) and L=2,3,4 (from simulation)
(Demand = 100 per period)

		Var (Q)			
Yield Distribution	Service Level	L = 1	L = 2	L = 3	L=4
$N(0.8, 0.05^2)$	80%	68.47	64.26	64.10	65.47
	90%	73.04	67.65	69.28	70.89
$N(0.7, 0.1^2)$	80%	560.36	520.84	543.31	446.59
	90%	678.20	604.27	592.58	501.08
$N(0.6, 0.1^2)$	80%	1114.70	1019.59	1051.62	849.23
	90%	1415.03	1217.96	1164.06	971.31
Beta (8,2)	80%	484.52	440.49	420.43	398.39
	90%	638.04	520.10	460.98	453.63
Beta (7,2)	80%	680.80	624.72	723.51	554.19
	90%	947.63	748.32	854.71	646.08
Beta (7,3)	80%	1271.99	1185.60	1006.45	833.74
	90%	1902.01	1533.33	1143.04	961.05

APPENDIX

Proof of Convergence of Moments of I

Proposition 1: $E(I_n) \dashrightarrow E(I)$ as $n \dashrightarrow \infty$ if $aE(P) > 1$.

Proof: Using the facts that $I_n = (1-ap_n)(I_{n-1} - D)$, I_{n-1} and p_n are independent, and the distribution of P is known, we have

$$E(I_n) = [1 - aE(P)]\,[E(I_{n-1}) - D].$$

We will prove that $E(I_n) \dashrightarrow E(I)$ as $n \dashrightarrow \infty$ by showing that

$$E(I_{n+1}) - E(I_n) < E(I_n) - E(I_{n-1}). \qquad \text{(A-1)}$$

for all $n > 3$. Now

$$E(I_n) - E(I_{n-1}) = D[aE(P) - 1] - aE(P)E(I_{n-1}).$$

With a little algebra, it can be shown that $E(I_{n+1}) - E(I_n) < E(I_n) - E(I_{n-1})$ if

$$E(I_{n-1}) < E(I_n),$$

or equivalently, if

$$E(I_{n-1}) < D[aE(P) - 1]/aE(P).$$

(A-2)

But if (A-2) holds, using the fact that $aE(P) > 1$, it is easy to show that

$$E(I_n) < D[aE(P) - 1]/aE(P)$$

and therefore by induction, that $E(I_n) < D[aE(P) - 1]/aE(P)$ for all n greater than the smallest value of n for which the inequality holds. Thus, we only need to show that (A-2) is satisfied for some finite value of n. Now

$$E(I_1) = D[aE(P) - 1] > D[aE(P) - 1]/aE(P)$$

so (A-2) is not satisfied for $n = 2$. However,

$$E(I_2) = D[aE(P) - 1]\,[2 - aE(P)] < D[aE(P) - 1]/aE(P)$$

if $aE(P) > 1$. Q.E.D.

Proposition 2: If $aE(P^2) - 2E(P) < 0$, then $E(I_n^2) \rightarrow E(I^2)$ as $n \rightarrow \infty$.

Proof: The proof parallels that of Proposition 1. Using a similar approach, we have

$$E(I_n^2) = [1 - 2aE(P) + a^2E(P^2)]\,[E(I_{n-1}^2) - 2DE(I_{n-1}) + D^2].$$

Thus, we have

$$E(I_{n+1}^2) - E(I_n^2) = [1 - 2aE(P) + a^2E(P^2)]\,[E(I_n^2) - E(I_{n-1}^2)]$$
$$+ [1 - 2aE(P) + a^2E(P^2)]\,\{2D\,[E(I_{n-1}) - E(I_n)]\}. \qquad \text{(A-3)}$$

For n sufficiently large, the second term on the right hand side of (A-3) goes to zero (as shown above), so our condition for convergence reduces to $1 - 2aE(P) + a^2E(P^2) < 1$, or

$$aE(P^2) - 2E(P) < 0. \qquad \text{Q.E.D.}$$

This condition places some restrictions on the characteristics of the yield distributions and the values of a that can be analyzed, but it is still fairly mild. For example, a distribution with $P\{p=0\} = 0.5$ and $P\{p=1\} = 0.5$ would be problematic, but for $f(p) = 12\,(p-0.5)^2$ which is a (bimodal, convex) quadratic density on [0,1] which is centered at 0.5, convergence is guaranteed for $a < 2.5$, and with a uniform distribution on [0,1], convergence is guaranteed for $a < 3$. It is also interesting to note that the condition is the same as that required for Var (I) to be positive.

Proposition 3: If $E\{(1-aP)^k\} < 1$ and if the lower moments converge, then $E(I_n^k) \rightarrow E(I^k)$.

Proof: The proof parallels that of Proposition 2, so we omit the details.