QUOTING CUSTOMER LEAD TIMES

IZAK DUENYAS
Department of Industrial & Operations Engineering
University of Michigan
Ann Arbor, MI 48109-2117

WALLACE J. HOPP
Department of Industrial Engineering
and Management Sciences
Northwestern University
Evanston, IL 60208

Technical Report 91-24

September 1991

QUOTING CUSTOMER LEAD TIMES

IZAK DUENYAS
Department of Industrial and Operations Engineering
The University of Michigan
Ann Arbor, Michigan 48109

WALLACE J. HOPP
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston Illinois 60208

Abstract

We consider the problem of quoting customer lead times in a manufacturing environment under a variety of modeling assumptions. First, we examine the case where capacity is effectively infinite relative to demand. For this case, we derive a closed-form expression for the optimal lead time quote and structural results under the assumption that price is fixed, so the firm competes on the basis of lead time alone, and under the assumption that the firm can choose both price and lead time. Second, we consider the case where capacity is finite and the firm processes jobs in first-come-first-served (FCFS) order. We prove optimality of different forms of control limit policies for the situations where the lead time is essentially dictated by the market and where firms are able to compete on the basis of lead time. Finally, we consider the case where the firm may choose to schedule jobs in other than FCFS order and give conditions under which the optimal due-date-quoting/order-scheduling policy will process jobs in earliest due date (EDD) order.

1 Introduction

In an increasingly global marketplace, manufacturing firms are being forced to compete on an expanding set of criteria. A recent practitioner-oriented publication summarized this historical trend as:

> *"How to do more" was emphasized in the 60s. "How to do it cheaper" became important in the 70s. "How to do it better" was certainly the theme of the 80s. But "How to do it quicker" will be key in the 90s (Charney 1991, p 1).*

The flood of practitioner literature focusing on lead times and customer responsiveness (see e.g., Blackburn 1990, Schmenner 1988, Stalk and Hout 1990, Thomas 1990, 1991); clearly support Charney's observation; speed is on the rise as a strategic competitive weapon.

In support of the growing interest on the part of practitioners for greater customer responsiveness, management science researchers have begun to establish a literature devoted to the analysis of lead times (see Karmarkar 1989 for a comprehensive review). A significant

amount of research has been devoted explicitly to cycle time reduction (see e.g., Baker 1987, Bechte 1982, Calabrese 1988, Dobson, Karmarkar and Rummel 1987, Dobson, Karmarkar and Rummel 1988, Hopp, Spearman and Woodruff 1990).

Other analytical research has focused on understanding lead times and their role in manufacturing systems. For instance, some researchers have concentrated on predicting the manufacturing lead time in systems (see e.g., Morton and Vepsalainen 1987, Ornek and Collier 1988, Shanthikumar and Sumita 1988). Others have worked to determine lead times within the manufacturing system (i.e., manufacturing or purchasing lead times, rather than customer lead times) that will improve system performance (see e.g., Hopp and Spearman 1989, Yano 1987a, Yano 1987b). Still others have examined the relationship between lead times and other parameters in manufacturing systems, such as lot sizes, inventories, dispatching rules, and customer priorities (see e.g., Eppen and Martin 1988, Karmarkar 1987, Karmarkar et al. 1985, Kekre and Udayabhanu 1988, Philiproom et al. 1987, Vepsalainen and Morton 1988). Finally, a significant number of researchers have considered the due-date-setting problem (see e.g., Baker 1984, Baker and Bertrand 1981, 1982, Seidmann and Smith 1981, Shanthikumar and Sumita 1988, Wein 1991).

Virtually all of the due-date-setting models consider the problem entirely from the perspective of the manufacturing firm. Due dates are selected to minimize holding costs, tardiness, fraction of late jobs, etc., in a variety of manufacturing contexts. However, to date very little work has focused on the customer perspective. As firms increasingly compete on the basis of lead time, the due date quoted to a potential customer will have a strong effect on whether he/she actually places an order. Under these conditions, lead times can have a strong effect on revenues as well as costs. As Karmarkar (1989, p 6) put it,

> *Presumably, lead times must be inversely related to market share or price premiums, or both (i.e., to total revenue). Certainly, managers are observed in practice to act as though shorter lead times confer a competitive advantage ...*

In this context, models that clarify the relationship between lead times, customer demand, and profitability, offer the potential to refine the use of lead time quoting as a strategic weapon.

Because the status of modeling research on the problem of setting lead times where demand is sensitive to the quoted lead times is still in its infancy, this paper is intended primarily as an impetus to further research. We restrict our attention to the situation where the firm quotes due dates to each customer independently. Within this scope, we concentrate on (1) modeling the appropriate costs and revenues under a variety of modeling assumptions and (2) characterizing the structure of the optimal policies. Our intent is to establish a modeling framework in which to consider lead times and to generate insight into the manner in which dynamically quoted lead times should incorporate the status of the manufacturing facility. Our hope is that future research will extend this framework to generate practical tools for assisting the lead time quoting process.

The remainder of the paper is organized as follows. Section 2 considers the case where capacity is effectively infinite relative to demand. Section 3 considers the case where capacity is finite and all jobs are processed in FCFS order. Section 4 considers the case where the firm can violate the FCFS order and addresses the resulting scheduling problem. The paper concludes in Section 5.

2 Infinite Capacity Case

We begin by considering the simplest case where plant capacity is large enough relative to demand to be reasonably considered infinite and where all customers are identical. This allows us to model the production environment as a $G/G/\infty$ queue. Customers arrive to the system according to some general distribution, and request one unit of product. The proportion of customers that actually place an order depends on the quoted lead time. Specifically, we suppose that a customer promised delivery in a time units has a probability of placing an order $p(a)$. (We assume that $p(a)$ is continuous and decreasing in a.) Each order generates net revenue (price minus production cost) of R. However, if the order is not filled on time, then the firm incurs a penalty. We let $F(\cdot)$ represent the distribution of production time and assume it to be continous with derivatives $f(\cdot)$ and $f'(\cdot)$.

2.1 Fixed Penalty

We first model the case where failure to deliver on time results in loss of the order. We model this by assuming that if production lasts more than a units of time, where a is the quoted lead time, then the firm incurs penalty C, where $C \geq R$. Letting π be the expected profit from a customer, we can express the problem of choosing an optimal lead time to quote as the following optimization problem:

$$\pi = \max_a p(a)(R - C(1 - F(a))) \tag{2.1}$$

We differentiate (2.1) twice with respect to a to get the following optimality conditions:

$$\frac{\partial \pi}{\partial a} = p'(a)(R - C(1 - F(a))) + p(a)Cf(a) = 0 \tag{2.2}$$

$$\frac{\partial^2 \pi}{\partial a^2} = p''(a)(R - C(1 - F(a))) + p'(a)Cf(a) + p(a)Cf'(a) < 0 \tag{2.3}$$

Hence, to calculate the optimal lead time, a^*, we must solve (2.2) for a subject to the second order condition in (2.3). In practice, solving (2.2) by numerical methods is straightforward. More interesting is the question of what insights can be gained from this model. Toward this end, we now solve (2.2) under the assumption that production takes an exponential amount of time with mean $b = 1/\mu$ and $p(a)$ is of the form $p(a) = e^{-\lambda a}$. (In Section 3, we will relax this assumption about the form of $p(a)$).

While these assumptions are intended primarily to make the model tractable, they are not grossly unrealistic. Highly variable production times, possibly approaching exponential, could result from the usual sources of manufacturing variability (e.g., machine failures) or from the fact that the product is really composed of multiple types, say standard or custom, requiring greatly different processing times.

The form we have chosen for $p(a)$ is more speculative. At this time, the authors are not aware of any empirical studies that will aid in setting $p(a)$. However, as mentioned previously, the strategic importance of reducing lead times has been stressed in many contexts. The form of $p(a)$ assumed in this section implies the importance of being competitive in lead times since this form implies a high demand for short lead times with demand falling off sharply as the lead time increases. Under the above assumptions, we can state the following

Theorem 1 *If* $p(a) = e^{-\lambda a}$, *and* $F(a) = 1 - e^{-\mu a}$, *then*

$$a^* = 1/\mu \ln(C(\lambda + \mu)/R\lambda) \tag{2.4}$$

Proof: The proof follows directly from conditions (2.2) and (2.3). □

This closed-form solution allows us to consider the effects of changing the parameters in our model on the optimal profit and lead time. We define $l = 1/\lambda$ and note that l can be interpreted as the mean lead time customers are willing to accept.

Theorem 2 *If* $F(a) = 1 - e^{-\mu a}$ *and* $p(a) = e^{-\lambda(a)}$, *then* $\frac{\partial \pi^*}{\partial R} \geq 0$, $\frac{\partial \pi^*}{\partial C} \leq 0$, $\frac{\partial \pi^*}{\partial l} \geq 0$, $\frac{\partial \pi^*}{\partial b} \leq 0$, $\frac{\partial a^*}{\partial R} \leq 0$, $\frac{\partial a^*}{\partial C} \geq 0$, $\frac{\partial a^*}{\partial l} \geq 0$, $\frac{\partial a^*}{\partial b} \geq 0$.

Proof: The proof is omitted. □

The results of Theorem 2 are fairly intuitive. Increasing unit revenue causes profits to increase. Similarly, increasing the penalty for failure to meet leadtime causes profits to decrease. When revenue is increased, the firm earns more from each customer whose order is delivered on time and hence decreases lead time to attract more customers. Not surprisingly, the firm responds to an increase in l by quoting longer lead times. This enables the firm to increase profit. Similarly, an increase in the mean processing time means that the firm will spend more time processing orders, which causes the firm to quote longer lead times and degrades profitability.

2.2 Variable Penalty

The above results assume that failure to deliver on time causes the firm to lose the order, effectively resulting in a fixed penalty. In many practical situations however, customers will not necessarily cancel their orders upon late delivery and hence the penalty to the firm depends on the amount of the delay. To address this case, we now assume that if the firm is late by x units of time, it incurs a penalty of $c(x)$. We further assume that $c(x)$ is increasing in x, and that $c(0) = 0$. In this case, the profit maximizing lead time quote is computed by solving

$$\pi = \max_a p(a)(R - \int_a^\infty c(y-a)f(y)dy) \tag{2.5}$$

Differentiating (2.5) with respect to a, we obtain the following optimality conditions

$$\frac{\partial \pi}{\partial a} = p'(a)\left(R - \int_a^\infty c(y-a)f(y)dy\right) + p(a)\int_a^\infty c'(y-a)f(y)dy = 0. \tag{2.6}$$

$$\begin{aligned}\frac{\partial^2 \pi}{\partial a^2} &= p''(a)\left(R - \int_a^\infty c(y-a)f(y)dy\right) + \\ &\quad 2p'(a)\int_a^\infty c'(y-a)f(y)dy - p(a)\int_a^\infty c''(y-a)f(y)dy < 0.\end{aligned} \tag{2.7}$$

Given $p(a)$, $F(a)$, and $c(a)$, numerical methods can be utilized to find a^* by using (2.6), (2.7). However if we alter the conditions of Theorem 1, to include a linear penalty for delay, we get the following

Theorem 3 *If* $F(a) = 1 - e^{-\mu a}$, $p(a) = e^{-\lambda a}$ *and* $c(x) = cx$, *the optimal solution to (2.5) is*

$$a^* = 1/\mu \ln \frac{(\lambda + \mu)c}{\lambda R \mu} \tag{2.8}$$

Proof: Substituting the values for $F(a)$, $p(a)$, and $c(x)$ in (2.6) and solving for a^* gives the expression in (2.8). To show that this solution is the global maximum, we look at the second order-condition which becomes,

$$\frac{\partial^2 \pi}{\partial a^2} = \lambda^2 e^{-\lambda a} R - \frac{(\lambda + \mu)^2}{\mu} c e^{-(\lambda+\mu)a} < 0$$

We note that π is concave for $a < \hat{a} = 1/\mu \ln \frac{c(\lambda+\mu)^2}{\mu \lambda^2 R}$, and convex for $a > \hat{a}$. However, by (2.8), $a^* < \hat{a}$, so a^* is optimal. □

We can derive the same intuitive results of Theorem 2 for the model with a variable lateness cost.

Theorem 4 *If* $F(a) = 1 - e^{-\mu a}$, $p(a) = e^{-\lambda(a)}$, *and* $c(x) = cx$ *then* $\frac{\partial \pi^*}{\partial R} \geq 0$, $\frac{\partial \pi^*}{\partial c} \leq 0$, $\frac{\partial \pi^*}{\partial l} \geq 0$, $\frac{\partial \pi^*}{\partial b} \leq 0$, $\frac{\partial a^*}{\partial R} \leq 0$, $\frac{\partial a^*}{\partial c} \geq 0$, $\frac{\partial a^*}{\partial l} \geq 0$, $\frac{\partial a^*}{\partial b} \geq 0$.

Proof: The proof is omitted.

2.3 Setting Price and Lead Time

In the previous section, we made the assumption that demand was sensitive to lead time but not sensitive to price. Clearly, in many situations the time a customer is willing to wait will depend on the price he/she is charged. In this section, we assume that the probability that a customer places an order depends both on price and lead time and is denoted by $p(R, a)$. Under this scenario, the firm has to quote both price and lead time. We formulate the problem for the case where if the firm does not deliver on time, there is a fixed penalty $C > R$. Since, in this case R is a decision variable as well, we set $C = qR$ where $q > 1$, so that the penalty for a lost order will reflect the price. We present the fixed cost case because it is simpler, the case with a variable penalty is entirely similiar.

Under the above assumptions, the profit function can be written as:

$$\pi = \max_{a,R} p(R, a) R(1 - q(1 - F(a))) \tag{2.9}$$

We differentiate (2.9) with respect to both a and R to get the following first-order conditions:

$$\frac{\partial \pi}{\partial R} = (1 - q(1 - F(a)))(\frac{\partial p(R, a)}{\partial R} R + p(R, a)) = 0 \tag{2.10}$$

$$\frac{\partial \pi}{\partial a} = (1 - q(1 - F(a)))R \frac{\partial p(R, a)}{\partial a} + p(R, a) R q f(a) = 0 \tag{2.11}$$

The second order conditions are:

$$(\frac{\partial^2 \pi}{\partial a \partial R})^2 - \frac{\partial^2 \pi}{\partial R^2} \frac{\partial^2 \pi}{\partial a^2} < 0; \quad \frac{\partial^2 \pi}{\partial R^2} < 0. \tag{2.12}$$

where the second-order partials can be calculated by differentiating (2.10) and (2.11) with respect to a and R.

We can solve for the optimal price and lead time by using numerical methods given any processing time distribution and function $p(R,a)$. However, to gain more insight into our model we now assume, as in the previous section, that the processing time distribution is exponential. In the previous section, where price was fixed, we suggested using $p(a) = e^{-\lambda a}$ for $p(a)$. The analagous choice for this case would be $p(R,a) = e^{-\lambda a R}$. However, this form assumes that as long as a given increase in either price or lead time is matched by a proportional decrease in the second variable, demand remains the same. For example, given a price and lead time level, if we increase price to twice its current level, and decrease lead time to half its current level, the proportion of customers who decide to place orders would not change. Because this may be an unrealistic restriction, we generalize the model by letting $p(R,a) = e^{-\lambda R^n a}$, where $n > 0$.

Theorem 5 *If $F(a) = e^{-\mu a}$ and $p(R,a) = e^{-\lambda R^n a}$ where $n > 0$, then the optimal lead time a^* is the unique positive solution to*

$$e^{-\mu a}(\mu a n + 1) = \frac{1}{q} \tag{2.13}$$

The optimal price R^* satisfies:

$$R^n = 1/(\lambda a n) \tag{2.14}$$

Proof: Substituting the expressions for $p(R,a)$ and $F(a)$ in (2.10), we get

$$\frac{\partial \pi}{\partial R} = (1 - qe^{-\mu a})(e^{-\lambda R^n a})(1 - R^n \lambda a n) = 0 \tag{2.15}$$

from which we get (2.14). Similiar substitutions into (2.11) result in

$$\frac{\partial \pi}{\partial a} = e^{-\lambda R^n a} R(-\lambda R^n (1 - qe^{-\mu a}) + \mu q e^{-\mu a}) = 0 \tag{2.16}$$

Substituting (2.14) in (2.16) and rearranging terms we get (2.13).

To show that there is a unique value of a that satisfies (2.13), we express (2.13) in the form $w(a) = 1/q$ where $w(a) = e^{-\mu a}(\mu a n + 1)$. We note that $w(0) = 1$ and that $1/q < 1$. Differentiating $w(a)$ with respect to a, we get

$$w'(a) = \mu e^{-\mu a}(-\mu a n - 1 + n)$$

Hence, if $n \leq 1$, $w(a)$ is decreasing in a for all a, and therefore there is a unique a. If $n > 1$ then, $w(a)$ is increasing for $a < (n-1)/n\mu$, and decreasing for $a > (n-1)/n\mu$. Hence $a^* > (n-1)/n\mu$, and there can be only one positive solution to (2.13).

Checking the second-order conditions evaluated at R^* and a^* we find that

$$\frac{\partial^2 \pi}{\partial R^2} = (1 - q\bar{e}^{-\mu a})e^{-\lambda R^n a}(-n)/R < 0$$

and after some algebra, we find that

$$\left(\frac{\partial^2 \pi}{\partial R \partial a}\right)^2 - \frac{\partial^2 \pi}{\partial R^2}\frac{\partial^2 \pi}{\partial a^2} = e^{-2\lambda R^n a}\frac{\mu^2 n}{\mu a n + 1}\left(-1 + \frac{n}{\mu a n + 1}\right).$$

But, we have already proved above that the unique positive solution a^* satisfies, $n - 1 - \mu a^* n < 0$, so $n/(\mu a^* n + 1) < 1$. Hence the solution, a^* and R^*, defined by (2.13) and (2.14) is the global maximum. □

This closed-form solution permits us to derive some structural results for the case where the firm chooses both price and lead time.

Theorem 6

$$\frac{\partial a^*}{\partial \mu} \leq 0; \quad \frac{\partial R^*}{\partial \mu} \geq 0; \quad \frac{\partial a^*}{\partial q} \geq 0; \quad \frac{\partial R^*}{\partial q} \leq 0.$$

Proof: The proof is omitted.

Theorem 6 states that as the firm's production process becomes faster, it becomes optimal to quote lower lead times and charge more for each unit. Hence, the benefits of a faster production process are two-fold. The firm can quote lower lead times thus increasing its demand and also charge more for each order, thus increasing its profits. In contrast, as the relative penalty for each unit increases, it becomes optimal to increase lead times and to lower prices.

The analysis and results are very similiar for the case where the penalty for failure to deliver on time is proportional to the amount of tardiness.

3 Finite Capacity Case

In the previous section, price and lead time were set under the assumption of infinite capacity. In this section, we consider the case where the plant has finite capacity. Because this line of research is still in its early stages and we are searching for qualitative insights rather than quantitative exactness, we will assume for the sake of tractability that the plant can be modelled as an $M/M/1$ queue. Customers arrive to the system according to a Poisson process with rate λ. They are quoted a lead time, which may depend on the system load, and decide whether or not to place an order. Once an order is placed, we assume it is never cancelled and that orders are filled one at a time according to an exponential distribution with rate μ. We first treat the case where the market essentially dictates the acceptable lead time. Then, we consider the case where the choice of leadtime is left to the firm and the customer's probability of placing an order depends on the quoted lead time. Throughout this section, we assume that all customer orders are handled on a first-come, first-served basis. In Section 4, we consider the possibility that the firm may choose not to follow a FCFS service discipline.

3.1 Industry Standard for Lead Times

In this section, we assume that the firm is in a market where both the price and acceptable leadtime are fixed. This would represent the case where customers expect a certain delivery speed but do not benefit from shorter delivery times. Specifically, we assume that the probability a customer who is quoted a lead time of a places an order is of the form:

$$p(a) = \begin{cases} 1 & \text{if } a \leq \hat{a} \\ 0 & \text{otherwise} \end{cases}$$

If the market is competitive and the customer expectations are economically achievable, then the market equilibrium could result in a situation where all firms offer the same price and lead time. We refer to this market determined lead time, $\hat{a}$, as the *industry standard* lead time.

Under these conditions, the company's only control over the customers is to accept or reject them. Each time a new customer arrives to the system, the company has the option to accept his/her order by quoting a lead time of $\hat{a}$, or rejecting it by quoting a higher lead time. If an order is accepted, and it is late by x units of time, we assume that a penalty of cx is incurred by the firm. We can formulate the problem of optimally quoting lead times as a Semi-Markov Decision Process (SMDP) where decisions are made at customer arrival times and the objective is to maximize average profits.

We define the states $k \geq 0$ as the number of orders in the system. We let v_k be the relative value function of being in state k, with $v_0 = 0$, and let g be the average profit per period (i.e., profit per arrival). At each decision epoch, the firm has the choice to accept or reject the new order. (Note that the periods are exponentially distributed with parameter λ, independent of whether or not the customer is accepted. Hence, there is no need to express g as profit per unit time as is normally done in general SMDP's). Letting q_i be the probability that i customers are served between any two arrivals, we can write the SMDP as

$$g + v_k = \max\{R - \varphi_{k+1} + w_{k+1}, w_k\} \tag{3.17}$$

where

$$\varphi_k = \int_{\hat{a}}^{\infty} c(x - \hat{a}) f_k(x)$$

and f_k is the density of the Erlang-k distribution, and

$$w_k = \sum_{i=0}^{k-1} q_i v_{k-i} + v_0 \sum_{i=k}^{\infty} q_k$$

We denote the optimal action in state k by a_k^*. The possibilities for a_k^* are $\hat{a}$ and ∞, where ∞ denotes the "reject" option.

We can use this SMDP formulation to show that the optimal policy has a control-limit form, such that the optimal decision is to admit customers when the number of customers is less than k^*, and to reject them otherwise. To do this, we begin with the following lemma.

Lemma 1 *If $w_{k+2} - 2w_{k+1} + w_k \leq \varphi_{k+2} - \varphi_{k+1}$, for all k, then there exists k^* such that $a_k^* = \hat{a}$ for all $k < k^*$ and $a_k^* = \infty$ for all $k \geq k^*$.*

Proof: Suppose the optimal action in state k is to reject the customer. Then from (3.17), this implies $w_k > w_{k+1} + R - \varphi_{k+1}$. The condition of the lemma implies $-w_{k+2} + 2w_{k+1} - w_k \geq -\varphi_{k+2} + \varphi_{k+1}$. Adding the two inequalities yields $w_{k+1} > w_{k+2} + R - \varphi_{k+2}$, which by (3.17) implies that the optimal action in state $k+1$ is to reject as well. The result follows from this. □.

Lemma 1 gives a sufficient condition for the optimal policy to have a control limit structure. However, this condition is not in terms of basic problem parameters. The following lemma and theorem extend this result to give us the desired condition.

Lemma 2 *$w_{k+2} - 2w_{k+1} + w_k \leq \varphi_{k+2} - \varphi_{k+1}$ for all k if and only if $v_{k+2} - 2v_{k+1} + v_k \leq \varphi_{k+2} - \varphi_{k+1}$ for all k.*

Proof: Suppose that for all k, $v_{k+2} - 2v_{k+1} + v_k \leq \varphi_{k+2} - \varphi_{k+1}$. Then, we can write,

$$w_{k+2} - 2w_{k+1} + w_k = \sum_{i=0}^{k} q_i(v_{k+2-i} - 2v_{k+1-i} + v_{k-i}) + q_{k+1}(v_1)$$

It is simple to show using a straightforward recursive argument that $v_k \leq 0$, for all k. Hence, we can write

$$w_{k+2} - 2w_{k+1} + w_k \leq \sum_{i=0}^{k} q_i(\varphi_{k+2-i} - \varphi_{k+1-i})$$

It is also easy to show that $(\varphi_{k+2} - \varphi_{k+1})$ is increasing in k. Hence, we have

$$\sum_{i=0}^{k} q_i(\varphi_{k+2-i} - \varphi_{k+1-i}) \leq \varphi_{k+2} - \varphi_{k+1}$$

and therefore the inequality holds for w_k.

To prove the reverse, we note that by Lemma 1, we know that if $w_{k+2} - 2w_{k+1} + w_k \leq \varphi_{k+2} - \varphi_{k+1}$, the optimal policy has a control-limit structure. There are four possibilities for the sequence of optimal actions in states k through $k+2$ that are consistent with this structure. Expressing the actions for states k, $k+1$, $k+2$ in order, these are $(\hat{a},\hat{a},\hat{a})$, $(\hat{a},\hat{a},\infty)$, $(\hat{a},\infty,\infty)$, and (∞,∞,∞).

If $(\hat{a},\hat{a},\hat{a})$ is chosen, then we have

$$v_{k+2} - 2v_{k+1} + v_k = w_{k+3} + w_{k+1} - 2w_{k+2} + 2\varphi_{k+2} - \varphi_{k+3} - \varphi_{k+1} \leq \varphi_{k+2} - \varphi_{k+1}$$

It is simple to show that cases $(\hat{a},\hat{a},\infty)$, $(\hat{a},\infty,\infty)$, and (∞,∞,∞) also yield the inequality

$$v_{k+2} - 2v_{k+1} + v_k \leq \varphi_{k+2} - \varphi_{k+1}$$

and hence the lemma is proven. □

Theorem 7 *The optimal policy in SMDP (3.17) is of the form*

$$a_k^* = \begin{cases} \hat{a} & \text{if } k < k^* \\ \infty & \text{if } k \geq k^* \end{cases}$$

Proof: We prove the result by a convergence argument. Let $w_k^0 = 0$, and $g + v_k^{i+1} = \max\{R - \varphi_{k+1} + w_{k+1}^i, w_k^i\}$ for all i. Then $v_k^1 = -g + \max\{R - \varphi_{k+1}, 0\}$, and $v_{k+2}^1 - 2v_{k+1}^1 + v_k^1 \leq \varphi_{k+2} - \varphi_{k+1}$ for all k. Furthermore, for all i and k, if $v_{k+2}^i - 2v_{k+1}^i + v_k^i \leq \varphi_{k+2} - \varphi_{k+1}$, then $w_{k+2}^i - 2w_{k+1}^i + w_k^i \leq \varphi_{k+2} - \varphi_{k+1}$ by Lemma 2. Since $w_{k+2}^i - 2w_{k+1}^i + w_k^i \leq \varphi_{k+2} - \varphi_{k+1}$, for all i by Lemma 1, the optimal policy is of control-limit form. Since state 0 is accessible from every state, it follows that $v_k^i \to v_k$, and $w_k^i \to w_k$ as $i \to \infty$ (Ross 1983), and the result is proven. □

Holding Costs

The reason that the solution to (3.17) can be characterized as a simple control limit policy is that the model provides no incentive for quoting lead times that are lower than the industry standard. The only costs in the model are those resulting from late delivery. There is no penalty for finishing an order before its due date. In practice, however, the firm may not be able to ship orders early. A firm whose production times are much shorter than the industry standard leadtime may incur a large finished goods carrying cost if it quotes industry standard lead times. In this case it may actually be attractive to quote lead times that are below the industry standard, even if doing so does not increase customer demand.

To model this scenario, we let h be the holding cost per unit time of WIP and define the cost of quoting lead time a with k customers in the system as

$$\varphi_k(a) = ha + \int_a^\infty c(y-a) f_k(y) dy \tag{3.18}$$

Our modelling assumptions here are:

1. Holding costs begin to accrue at the time an order is placed (e.g., the firm immediately purchases raw materials).
2. The holding cost is constant over the production cycle (e.g., we ignore "added value" issues by treating labor and capital as fixed costs).
3. The penalty for being late is propotional to the amount of tardiness.
4. The holding cost continues to accrue until the order is shipped. We assume that c includes the holding cost h, and therefore $c > h$. We ignore the time lag between the ship date and the date of customer payment since this lag will typically result in a constant amount of increase in holding costs, which is independent of the due date quotes.

As in the previous formulation, we let $\hat{a}$ be the industry standard lead time. Under these assumptions, we can formulate the average-cost SMDP as:

$$g + v_k = \max\{\max_{a \le \hat{a}} R - \varphi_{k+1}(a) + w_{k+1}; w_k\} \tag{3.19}$$

Note that because the holding cost may provide incentive to quote lead times below the industry standard, SMDP (3.19) requires a larger set of decision variables than SMDP (3.17). The possibilities for a_k^*, the optimal decision in state k, are $a \le \hat{a}$ and ∞.

Under these conditions, we can again show that a control-limit policy is optimal for (3.19) but, as one would expect, the optimal lead times are state dependent. To prove this result, we define a_k as $a_k = \arg\min_{a \le \hat{a}} \varphi_k(a)$ (i.e., a_k represents the lead time that minimizes the holding plus tardiness cost for a customer that arrives to see k customers already in the system). We also require the following technical lemmas:

Lemma 3 *If* $w_{k+2} - 2w_{k+1} + w_k \le \varphi_{k+2}(a_{k+2}) - \varphi_{k+1}(a_{k+1})$ *then there exists* k^* *such that* $0 \le a_k^* \le \hat{a}$ *for* $k < k^*$ *and* $a_k^* = \infty$ *for* $k \ge k^*$.

Proof: The proof is similiar to that of Lemma 1.

Lemma 4 *$\varphi_k(a_k)$ is convex in k.*

Proof: It is enough to show that for any x, y such that $x \leq \hat{a}$, and $y \leq \hat{a}$,

$$\varphi_{k+2}(y) + \varphi_k(x) \geq 2\varphi_{k+1}((x+y)/2) \tag{3.20}$$

since by letting $y = a_{k+2}$ and $x = a_k$, we have

$$\varphi_{k+2}(a_{k+2}) + \varphi_k(a_k) \geq 2\varphi_{k+1}((a_{k+2} + a_k)/2) \geq 2\varphi_{k+1}(a_{k+1})$$

If we let X_{k+1} denote the (random) amount of time to finish $k+1$ orders, then we can write X_k as $X_{k+1} - Z$, and X_{k+2} as $X_{k+1} + Z$ where Z is the random amount of time to finish one order. Then we can express the left side of (3.20) as

$$\varphi_{k+2}(y) + \varphi_k(x) = hy + hx + cE[(X_{k+1} + Z - y)^+] + cE[(X_{k+1} - Z - x)^+]$$

Using the identity, $E[a^+] + E[b^+] \geq E[(a+b)^+]$, we obtain (3.20) and the proof is complete. □.

We can now prove a revised control-limit result for the model with holding costs.

Theorem 8 *There exists a number k^* such that the optimal solution to SMDP (3.19) is of the form*

$$a_k^* = \begin{cases} \min\{\hat{a}, a_{k+1}\} & \text{if } k < k^* \\ \infty & \text{if } k \geq k^* \end{cases}$$

where a_k is the solution to

$$F_k(a_k) = \frac{c-h}{c}. \tag{3.21}$$

Proof. The proof is similiar to that of Theorem 1. Using Lemmas 3 and 4, we can show that $w_{k+2} - 2w_{k+1} + w_k \leq \varphi_{k+2}(a_{k+2}) - \varphi_{k+1}(a_{k+1})$ if and only if $v_{k+2} - 2v_{k+1} + v_k \leq \varphi_{k+2}(a_{k+2}) - \varphi_{k+1}(a_{k+1})$. A convergence argument proves the control limit result. Expression (3.21) is derived by taking the derivative of $\varphi_k(a)$ in (3.18) and setting it equal to 0. □.

3.2 Variable Leadtimes

In this section, we assume that there is no industry standard for lead times, and that firms compete on the market on the basis of lead times. As in Section 2, we assume that customers arrive to the system according to a Poisson process with rate λ. A customer who is quoted lead-time a places an order with probability $p(a)$ where $p(a)$ is decreasing in a. We further assume that there exists a finite $\hat{a}$ such that $p(a) = 0, a > \hat{a}$. Each order brings in a revenue of R units to the firm and there is a penalty of c per unit time per order for late orders. Customers are served in the order they arrive and have exponentially distributed service times with mean $1/\mu$.

Under these conditions, we can formulate the problem of quoting optimal lead times as a (SMDP) where the states are the number of customer orders in the system when a new customer comes in. Defining v_k and g and w_k and letting a_k^* denote the optimal action in state k, we can write the SMDP recursion as

$$g + v_k = \max\{\max_a p(a)(R - \varphi_{k+1}(a) + w_{k+1}); w_k\} \tag{3.22}$$

Since, we are interested in characterizing the optimal policy, we first note that there exists a state k^* such that for all $k \geq k^*$, the optimal decision is to reject the customer (i.e., $a_k^* = \infty$ for $k \geq k^*$). Obviously, an upper bound on k^* is $\inf_k : \varphi_k(\hat{a}) > R$, which is the point where the expected tardiness cost associated with a customer is larger than the revenue generated by that customer. This observation allows us to restrict our problem to a finite state space without loss of optimality.

Secondly, we note that the problem of choosing optimal lead times at the arrival points of customers is equivalent to the problem of choosing arrival rates to an M/M/1 queue. If we quote lead time a to a customer, then we are in effect setting the arrival rate to the queue of customers who have decided to place orders to be $\lambda p(a)$. In the problem where we choose arrival rate λ_k when there are k customers in the system, the state 0 is reachable under any stationary policy. Hence by Ross (1983), a stationary policy is optimal.

Let $\pi_\sigma(k,j)$ be the expected profit between one visit from state k to state j under stationary policy σ, and $t_\sigma(k,j)$ be the expected number of customers that arrive to the system (but do not necessarily place an order) until that visit to state j. If we let g_σ be be the long-run average profit under policy σ, then by renewal-reward theory, we have $g_\sigma = \pi_\sigma(k,k)/t_\sigma(k,k)$. Moreover, $\pi_\sigma(k,k) - g^* t_\sigma(k,k) \leq 0$, with equality obtained by maximizing $\pi_\sigma(k,k) - g^* t_\sigma(k,k)$ over all σ. However, Stidham and Weber (1989) have shown that for any g, the quantity $\pi_\sigma - g t_\sigma(k,k)$ equals the total g-revised profit obtained between any two visits to state k. That is, if we solve the problem of maximizing the expected profit between any two visits to state k, where we subtract g^* from the expected profit in each state, we have solved (3.22). Since 0 is a reachable state from any state, we consider the problem of optimally reaching state 0.

Letting $\pi(k,0)$ represent the optimal expected profit starting in state k until state 0 is reached for the first time, renewal reward theory allows us to write:

$$\pi(i,0) = \max_a \frac{-g^* + \lambda p(a)(R - \varphi_{k+1}(a) + \pi(k+1,0)) + \lambda(1-p(a))\pi(k,0) + \mu\pi(k-1,0)}{\lambda + \mu} \tag{3.23}$$

which after arranging terms and letting $\lambda(a) = \lambda p(a)$ becomes,

$$\pi(k,0) = \max_a \frac{-g^* + \lambda(a)(R - \varphi_{k+1}(a) + \pi(k+1,0)) + \mu\pi(k-1,0)}{\lambda(a) + \mu} \tag{3.24}$$

This can be further simplified to yield

$$\pi(k,0) = \max_a \frac{-g^* + \lambda(a)(R - \varphi_{k+1}(a))_+ \lambda(a)(\pi(k+1,0) - \pi(k,0))}{\mu} + \pi(k-1,0) \tag{3.25}$$

By the left-skip-free-property (Wijngaard and Stidham 1986) we know that

$$\pi(k,0) = \pi(k,k-1) + \pi(k-1,0) \tag{3.26}$$

Combining (3.25) and (3.26), and letting $B_{k+1}(a) = \lambda(a)(R - \varphi_{k+1}(a))$, we get

$$\pi(k,k-1) = \max_a \frac{-g^* + B_{k+1}(a) + \lambda(a)\pi(k+1,k)}{\mu} \tag{3.27}$$

With this we are able to prove the following technical lemma:

Lemma 5 *$\pi(k, k-1)$ is decreasing in k.*

Proof. We know that for all $k \geq k^*$, $\lambda(a) = 0$. Hence for all $k \geq k^*$, $\pi(k, k-1) = -g^*/\mu$. For $k < k^*$, we can write

$$\pi(k, k-1) = \frac{B_{k+1}(a_k^*) + \lambda(a_k^*)\pi(k+1, k) - g^*}{\mu}$$

Now, suppose that $\pi(j, j-1) \geq \pi(j+1, j)$ for $j = k, \ldots, k^*$. Then

$$\pi(k-1, k-2) = \max_a \frac{B_k(a) + \lambda(a)\pi(k, k-1) - g^*}{\mu} \geq \frac{B_k(a_k^*) + \lambda(a_k^*)\pi(k, k-1) - g^*}{\mu}$$

However, $B_k(a_k^*) \geq B_{k+1}(a_k^*)$ and by the induction assumption $\pi(k, k-1) \geq \pi(k+1, k)$. Hence, we have shown $\pi(k-1, k-2) \geq \pi(k, k-1)$ and the proof is complete. □

We can now show that the optimal policy for the problem of optimally bringing the system to state 0 is a monotonic policy, (i.e., it is optimal to have decreasing effective arrival rates $\lambda(a_k^*)$ by quoting longer lead times when there are more customers in the system).

Lemma 6 *The optimal solution to (3.23), has a_k^* increasing in k.*

Proof: From (3.26), we have $\pi(k, 0) = \pi(k, k-1) + \pi(k-1, 0)$ where

$$\pi(k, k-1) = \max_a \delta(k, a)$$

and

$$\delta(k, a) = \frac{-g^* + B_{k+1}(a) + \lambda(a)\pi(k+1, k)}{\mu}$$

Choose a_1 and a_2 such that $a_1 > a_2$. It is sufficient to show that $\delta(k, a_1) - \delta(k, a_2)$ is increasing in k.

$$\delta(k, a_1) - \delta(k, a_2) = \frac{B_{k+1}(a_1) - B_{k+1}(a_2)}{\mu} + \frac{(\lambda(a_1) - \lambda(a_2))\pi(k+1, k)}{\mu} \tag{3.28}$$

The second-term on the right-hand side of (3.28) is increasing in k, since $\lambda(a_1) - \lambda(a_2)$ is negative, and by the previous lemma $\pi(k+1, k)$ is decreasing in k. Also,

$$\begin{aligned} B_{k+2}(a_1) - B_{k+2}(a_2) - B_{k+1}(a_1) + B_{k+1}(a_2) = \lambda(a_2)(\varphi_{k+2}(a_2) - \\ \varphi_{k+1}(a_2)) - \lambda(a_1)(\varphi_{k+2}(a_1) - \varphi_{k+1}(a_1)) \geq 0 \end{aligned}$$

Hence, $\delta(k, a_1) - \delta(k, a_2)$ is increasing in k and the proof is complete. □.

Since we have shown above that the problem of maximizing expected profits until reaching state 0 is equivalent to problem (3.22), we have also proved the following

Theorem 9 *The optimal solution to (3.22) has a_k^* increasing in k.*

4 Finite Capacity with Scheduling Considerations

In the previous section, we assumed that each customer order is filled on a FCFS basis. In some situations this may be reasonable. For instance, if customers have information about the status of other orders and are displeased if they learn that a customer who ordered later was served earlier, the firm may choose to maintain a FCFS discipline.

However, in many cases the company is free to choose the sequence in which the orders are to be filled. One case where it might be adventageous to the firm not to fill orders on a FCFS basis is the situation where the production facility is prone to "lucky streaks." If such a streak occurs and several orders were finished much earlier than expected, there might be a great deal of slack in the due dates of the remaining orders. If a new customer arrives at this point, the firm may be wise to quote the customer a low lead time, thereby increasing the chances of getting the order. If placed, the order could then be placed at the begining of the queue without jeopordizing the integrity of the due dates of the existing orders.

A major difficulty of formulating the finite capacity problem with the option of servicing orders out of FCFS sequence is that each time a new customer arrives to the system, we need information on how much time is left until the due date of each order. This requires us to define states as $(k, t_1, t_2, \ldots, t_k)$, where k is the number of customer orders in the system at the time a new customer arrivs, and $t_j, j = 1, \ldots, k$, denotes the amount of time left until the due date of the j^{th} in the queue. (We take t_1 to be the due date of the order currently in service.) Note that t_j could be negative, in which case it denotes how much time has passed since the due date of the j^{th} order.

Letting g^* denote the optimal average profit per arrival as in the previous section, and letting $v(k, t_1, ., t_j, .., t_k)$ denote the relative value function, we can write the SMDP recursion as:

$$g^* + v(k, t_1, \ldots, t_j, \ldots, t_k) = \max_{a, j=1,\ldots,k+1} \{p(a)(R - \varphi_j(a) - \sum_{i=j}^{k} \varphi_{i+1}(t_i) - \varphi_i(t_i) + w(k+1, t_1, \ldots, t_{j-1}, a, t_j, \ldots, t_k)) + (1 - p(a))w(k, t_1, \ldots, t_k)\} \tag{4.29}$$

Note that in this case the firm chooses not only the lead time to quote but also where to place the new order if the customer decides to place an order. If the order is placed on the j^{th} position, then the expected amount of time that orders j through k are delayed increases. Hence, the expected costs include not only the delay cost of the new order but of the orders that are displaced as well. As in the previous section, the w values denote the expected profits after the new order is placed. However, in this case they are slightly more complicated. Let q_i denote the probability that there are i services until the next arrival, then

$$w(k, t_1, \ldots, t_k) = \sum_{i=0}^{k-1} q_i \int_{t=0}^{\infty} v(k - i, t_{i+1} - t, \ldots, t_k - t)\lambda e^{-\lambda t} dt + \sum_{i=k}^{\infty} q_i v(0). \tag{4.30}$$

As in the previous section, we seek structural results for the optimal policy. This requires the following technical lemmas:

Lemma 7 *For all $t > 0$ and for all j, k, and $t_1, \ldots, t_k$,*

$$v(k, t_1, \ldots, t_j, \ldots, t_k) \geq v(k, t_1, \ldots, t_j - t, \ldots, t_k)$$

Proof: The proof is omitted.

Lemma 8 *Let $t_1 \leq t_2 \leq \cdots \leq t_k$. If for all k, and j and*

1. *for $a < t_j$, we have*
   $w(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k) \geq w(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k)$,
2. *for $a > t_j$, we have*
   $w(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k) \geq w(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k)$.

then an optimal policy will sequence customer orders according to an earliest due date (EDD) protocol.

Remark: The condition states that if there is a single job that is out of EDD order, then any exchange with an adjacent job that brings it closer to EDD increases profits.

Proof: Suppose that there are k customers in the system at the time of an arrival and that they are in EDD order, i.e., $t_1 \leq t_2 \leq \cdots \leq t_k$. Consider a lead time quote a such that $t_j \leq a \leq t_{j+1}$ for some $j = 1, \ldots, k$. If the customer is placed in the $j+1^{st}$ (i.e., EDD) position, expected total profits are:

$$\begin{aligned} r_1 &= p(a)(R - \varphi_{j+1}(a) - \sum_{i=j+1}^{k} (\varphi_{i+1}(t_i) - \varphi_i(t_i)) + w(k+1, t_1, \ldots, t_j, a, t_{j+1}, \ldots, t_k) + \\ &\quad (1 - p(a))w(k, t_1, \ldots, t_k). \end{aligned} \tag{4.31}$$

Alternatively, if we place the customer in the z^{th} position where $z \leq j$, then expected total profits are:

$$\begin{aligned} r_2 &= p(a)(R - \varphi_z(a) - \sum_{i=z}^{k} (\varphi_{i+1}(t_i) - \varphi_i(t_i)) + w(k+1, t_1, \ldots, a, t_z, \ldots, t_{j+1}, \ldots, t_k) + \\ &\quad (1 - p(a))w(k, t_1, \ldots, t_k). \end{aligned} \tag{4.32}$$

Comparing (4.31) and (4.32), we find that

$$\begin{aligned} r_2 - r_1 &= p(a)((\varphi_{j+1}(a) - \varphi_z(a)) - (\sum_{i=z}^{j} (\varphi_{i+1}(t_i) - \varphi_i(t_i))) + \\ &\quad (1 - p(a))(w(k+1, t_1, \ldots, a, t_z, \ldots, t_{j+1}, \ldots, t_k) - \\ &\quad w(k+1, t_1, \ldots, t_j, a, t_{j+1}, \ldots, t_k)). \end{aligned} \tag{4.33}$$

By repeated application of assumption (2), it follows that $w(k+1, t_1, \ldots, a, t_z, \ldots, t_{j+1}, \ldots, t_k) \leq w(k+1, t_1, \ldots, t_j, a, t_{j+1}, \ldots, t_k)$. Now, notice that we can write $\varphi_{j+1}(a) - \varphi_z(a)$ as $\sum_{i=z}^{j} (\varphi_{i+1}(a) - \varphi_i(a))$. Hence we can write the first part of (4.33) as $p(a)(\sum_{i=z}^{j} ((\varphi_{i+1}(a) - \varphi_i(a)) - (\varphi_{i+1}(t_i) - \varphi_i(t_i)))$. But, by assumption, $a \geq t_j \geq t_{j-1} \geq \cdots \geq t_1$. Hence, $(\varphi_{i+1}(a) - \varphi_i(a)) \leq (\varphi_{i+1}(t_i) - \varphi_i(t_i))$ and we have shown $r_2 - r_1 \leq 0$, which proves that it is not optimal to place the order in any position less than the $j+1^{st}$ position.

To show that we would not place the new order in position $z > j+1$, we let r_3 denote the expected profit if we place the new order in position $z > j+1$. It is straightforward to show that

$$\begin{aligned} r_3 - r_1 &= p(a)((\varphi_{j+1}(a) - \varphi_z(a)) + \sum_{i=j+1}^{z-1} (\varphi_{i+1}(t_i) - \varphi_i(t_i))) + \\ &\quad (1-p(a))(w(k+1, t_1, \ldots, t_{j+1}, \ldots, a, t_z, \ldots, t_k) - \\ &\quad w(k+1, t_1, \ldots, a, t_{j+1}, \ldots, t_k)) \end{aligned} \tag{4.34}$$

Again, the second part of (4.34) is negative by repeated application of assumption (1). We can write the first part as $p(a)(\sum_{i=j+1}^{z-1} -(\varphi_{i+1}(a) - \varphi_i(a)) + (\varphi_{i+1}(t_i) - \varphi_i(t_i)))$. Since $a \le t_{j+1} \le t_{j+2} \le \cdots \le t_z$, for each $i \ge j+1$, we have $(\varphi_{i+1}(t_i) - \varphi_i(t_i)) \le (\varphi_{i+1}(a) - \varphi_i(a))$, and we have $r_3 - r_1 \le 0$.

Since the choice of a was arbitrary, the proof is complete. □

Lemma 9 *Let $t_1 \le t_2 \le \cdots \le t_k$. If for all k, j and*

1. *for $a < t_j$, we have*
   $v(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k) \ge v(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k)$,
2. *for $a > t_j$, we have*
   $v(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k) \ge v(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k)$

then

1. *for $a < t_j$, we have*

$$w(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k) \ge w(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k) \tag{4.35}$$

2. *for $a > t_j$, we have*

$$w(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k) \ge w(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k) \tag{4.36}$$

Proof: We prove the result for the first case, i.e., $a < t_j$. The case where $a > t_j$ is completely analogous. By (4.30), w is a combination of v's. We fix t, the time passed since the arrival of the last customer and m, the number of customers that have been served since then. If $m \ge j+1$, then both sides of (4.35) will equal $v(k+1-m, t_{m+1} - t, t_{m+2} - t, \ldots, t_k - t)$. If $m < j$, then the left hand side of (4.35) will be $v(k+1-m, t_{m+1} - t, \ldots, a - t, t_j - t, t_{j+1} - t, \ldots, t_k - t)$, while the right hand side will equal $v(k+1-m, t_{m+1} - t, \ldots, t_j - t, a - t, t_{j+1} - t, \ldots, t_k - t)$, and by assumption the left hand side is greater. Now, if $m = j$, then the left hand side will equal $v(k+1-m, t_j - t, t_{j+1} - t, \ldots, t_k - t)$ and the right-hand side will be $v(k+1-m, a - t, t_{j+1} - t, \ldots, t_k - t)$, and by Lemma 7, the left hand side is greater. Since we compared both sides for any value of t and m, unconditioning on these values will preserve the inequality. □.

Lemma 10 *Under the conditions of Lemma 8, and if a new job is not allowed to be placed in the j^{th} position unless all the jobs it displaces have time remaining until their due dates of at least $(i+1)/\mu$, $i = j, \ldots, k+1$, then*

1. *for* $a < t_j$, *we have*
   $v(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k) \geq v(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k)$,

2. *for* $a > t_j$, *we have*
   $v(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k) \geq v(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k)$

Proof: We prove the result for the case $a < t_j$, the case where $a > t_j$ is completely analogous.

We compare $r_1 = v(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k)$, with $r_2 = v(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k)$. To do this suppose that we fix the position of the new job and the leadtime that we are quoting to the new customer for both options. That is, suppose that we quote a lead time b, and place the customer on the y^{th} position for both states. Then, with probability $1 - p(b)$, the customer will not place an order and the relative value of future revenues starting in state $(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k)$ is $q_1 = w(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k)$ and starting in state $(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k)$ it will be $q_2 = w(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k)$. But, by assumption, $q_1 \geq q_2$.

On the other hand, if the customer does place an order, there are three different cases.

Case 1: Suppose $y > j+1$. Then, for each case there will be a fixed revenue R, the expected delay penalty for the new customer $\varphi_y(b)$, the expected costs due to displacement, $\sum_{i=y}^{k+1} \varphi_{i+1}(t_{i-1}) - \varphi_i(t_{i-1})$, and the relative value of future revenues. Starting in state $(k+1, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, t_k)$, total revenues will be $s_1 = R - \varphi_y(b) - \sum_{i=y}^{k+1}(\varphi_{i+1}(t_{i-1}) - \varphi_i(t_{i-1})) + w(k+2, t_1, \ldots, t_{j-1}, a, t_j, t_{j+1}, \ldots, b, t_y, \ldots, t_k)$, and starting in state $(k+1, t_1, \ldots, t_{j-1}, t_j, a, t_{j+1}, \ldots, t_k)$, they will be $s_2 = R - \varphi_y(b) - \sum_{i=y}^{k}(\varphi_{i+1}(t_{i-1}) - \varphi_i(t_{i-1})) + w(k+2, t_1, \ldots, t_{j-1}, t_j, a, \ldots, b, t_y, \ldots, t_k)$. Revenue, expected delay and displacement costs are the same for the two starting states, while the relative value of future revenues are higher in s_1, by assumption. Hence $s_1 \geq s_2$.

Case 2: Suppose $y = j+1$. Then, in a similiar manner to Case 1, we can write $s_1 - s_2 = w(k+2, t_1, \ldots, t_{j-1}, a, b, t_j, \ldots, t_k) - w(k+2, t_1, \ldots, t_{j-1}, t_j, b, a, \ldots, t_k) - (\varphi_{j+2}(t_j) - \varphi_{j+1}(t_j)) + (\varphi_{j+2}(a) - \varphi_{j+1}(a))$. Since $a < t_j$, we have $(\varphi_{j+2}(a) - \varphi_{j+1}(a)) \geq (\varphi_{j+2}(t_j) - \varphi_{j+1}(t_j))$. By assumption, $w(k+2, t_1, \ldots, t_{j-1}, a, b, t_j, \ldots, t_k) \geq w(k+2, t_1, \ldots, t_{j-1}, t_j, b, a, \ldots, t_k)$. Hence $s_1 - s_2 \geq 0$.

Case 3: Suppose $y \leq j$. Then, after some algebra we find that $s_1 - s_2 = w(k+2, t_1, \ldots, b, \ldots, a, t_j, t_{j+1}, \ldots, t_k) - w(k+2, t_1, \ldots, b, \ldots, t_j, a, t_{j+1}, \ldots, t_k) + (\varphi_{j+2}(a) - 2\varphi_{j+1}(a) + \varphi_j(a)) - (\varphi_{j+2}(t_j) - 2\varphi_{j+1}(t_j) + \varphi_j(t_j))$. Again, by the condition of the lemma, the difference of the two relative future value functions is positive. To show that the remaining terms are positive as well, let $g(x) = \varphi_{j+2}(x) - 2\varphi_{j+1}(x) + \varphi_j(x)$, and $F_j(x)$ be the convolution of j service times. Differentiating with respect to x, we find that $g'(x) = c(F_{j+2}(x) - 2F_{j+1}(x) + F_j(x))$. Since $F_j(x) = 1 - \sum_{i=0}^{j-1} \frac{e^{-\mu x}(\mu x)^j}{j!}$, we find that $g'(x) = ce^{-\mu x}(\frac{(\mu x)^j}{j!} - \frac{(\mu x)^{j+1}}{(j+1)!}) < 0$. Hence, $g(x)$ is decreasing for $x > (j+1)/\mu$. But, by assumption $a > (j+1)/\mu$, since the new job was placed before it, and $t_j > a$. Hence, again we have $s_1 \geq s_2$.

To complete the proof, we notice that $r_1 = \max_{b,y} p(b)s_1 + (1 - p(b))q_1$ and $r_2 = \max_{b,y} p(b)s_2 + (1 - p(b))q_2$. Since we have shown that $s_1 \geq s_2$ and $q_1 \geq q_2$ for all possible values for b and y, we conclude that $r_1 \geq r_2$. □.

Using the above three lemmas, we can prove the main result of this section:

Theorem 10 *If a new job is not allowed to be placed in the j^{th} position unless all the jobs it displaces have time remaining until their due dates of at least $(i+1)/\mu, i = j, \ldots, k+1$, then an optimal policy will process all jobs in EDD order.*

Proof: The proof is by induction. To begin the induction, let $w^1(k, t_1, \ldots, t_k) = 0$. Furthermore, let $v^i(k, t_1, \ldots, t_k) = \max_{a,j} p(a)(R - \varphi_j(a) + w^{i-1}(k+1, t_1, \ldots, a, t_j, \ldots, t_k)) + (1 - p(a))w^{i-1}(k, t_1, \ldots, t_k)$. For each i, the inequalities in Lemma 7, 8 and 9 are preserved, and the optimal schedule is an (EDD) schedule. Since the state 0 is reachable from any state, it is straightforward to show that $v^i(k, t_1, \ldots, t_k) \rightarrow v(k, t_1, \ldots, t_k)$ as $i \rightarrow \infty$. This completes the proof. □.

Theorem 10 gives us a condition under which jobs will be processed in EDD order. In essence, the condition that if a new job is not allowed to be placed in the j^{th} position unless all the jobs it is displaces have at least $(i+1)/\mu$ until their due dates is a serviceability condition. It precludes placing a new job ahead of other jobs if it causes their probability of being completed on time to fall below specified levels. Since the time to complete i jobs is distributed according to an Erlang-i distribution, these specified levels are given by the cdf of the Erlang evaluated at its mean. Hence, these levels range between 0.59 (for $i+1=2$) to 0.5 (as $i \rightarrow \infty$). Thus, a sufficient condition to ensure an EDD sequence is that no job can have its position preempted if such preemption would cause its probability of being completed on-time to fall below 59%. Since, in practice most firms have service targets of 90% or higher, this condition is not unrealistic.

The above result that, once due dates have been quoted, we will process jobs in EDD order certainly seems intuitive. Hence, it is interesting to note that this result is clearly dependent on the cost structure. If, for example, the penalty for a late job is fixed, then the EDD result does not hold. On the other hand, it is not difficult to show, using methods analagous to those above, that if the penalty for lateness is a quadratic or higher polynomial function of the lateness, then the EDD result holds without the serviceability condition of Theorem 10.

5 Conclusions and Further Work

In this paper, we have modeled the problem of quoting due dates in a manufacturing system under three levels of modeling assumptions.

At the simplest level, we assumed infinite capacity, so that optimal quotes did not depend on the current backlog. For this case, we were able to show that the optimal profits and quotes depended in an intuitive fashion on the problem parameters. We extended this model to allow the firm to quote both price and due date and, again, arrived at intuitive sensitivity results.

At the intermediate level, we restricted capacity and modeled the manufacturing system as a single server queue. Under exponential assumptions for customer arrivals and processing times, we considered two cases: (1) where the market dictates the acceptable (industry standard) lead time, so that the firm merely chooses whether or not to accept an order, and (2) where the firm is free to choose the lead time as well as whether or not to accpet the

order. In both cases, we demonstrated optimality of a control-limit policy. In second case, we showed that the optimal lead times are increasing in the number of orders in the work backlog.

At the most complex level, we considered the case where the firm has finite capacity and can choose the order in which to process jobs. In this case, it is quite possible that an optimal policy may call for processing orders in other than a FCFS sequence. We showed that, under a relatively mild serviceability condition, that the optimal lead-time-quoting/order-sequencing policy will result in jobs being processed according to an EDD sequence.

Much remains to be done to develop an effective arsenal of strategic lead time models. A set of research topics that seem promising are:

1. Developing an understanding of the dependence of customer demand on the quoted lead times. Empirical work on characterizing the $p(a)$ function is essential to making use of strategic models in practical settings.

2. Incorporating the long-term consequences of failing to deliver orders on time. Clearly, late orders affect a firms reputation and hence future demand. Empirical and modeling work are needed to address this issue.

3. Developing practical methods for attacking the combined lead time quoting/order sequencing problem. This is an extremely difficult problem. Structural results such as the sufficient condition for EDD sequences given in this paper can simplify the problem somewhat, but alternative modeling approaches, heuristics, and rules of thumb are needed before we can hope to offer practitioners much guidance in this area.

6 References

Baker, K., "Sequencing Rules and Due-Date Assignments in a Job Shop," Management Science **30**, (1984), 1093-2004.

Baker, K., "Lot Streaming to Reduce Cycle Time in a Flow Shop," The Amos Tuck School of Business Administration, Dartmouth College, Working Paper No. 203, (1987).

Baker, K., and J. Bertrand, "A Comparison of Due-Date Selection Rules," *AIIE Transactions* **13**, (1981), 123-131.

Baker, K., and J. Bertrand, "An Investigation of Due-Date Assignment Rules with Constrained Tightness," *Journal of Operations Management* **3**, (1982), 109-120.

Bechte, W., "Controlling Manufacturing Lead Time and Work-In-Process Inventory by Means of Load-Oriented Order Release," *Twenty-fifth Annual International Conference Proceedings of the American Production & Inventory Control Society*, APICS, Falls Church, VA, (1982), 67-72.

Bertrand, J., "The Effect of Workload Dependent Due-Dates on Job Shop Performance," *Management Science* **29**, (1983), 799-816.

Bertrand, J., "The Use of Workload Information to Control Job Lateness in Controlled and Uncontrolled Release Production Systems," *Journal of Operations Management* **3**, (1983), 67-78.

Blackburn, Joseph, *Time Based Competition*, Richard D. Irwin, Homewood, Illinois, (1990).

Calabrese, J., "Lot Sizing and Priority Sequencing to Minimize Manufacturing Cycle Time and Work-In-Process Inventory," working paper, Dept. of Industrial Engineering, Stanford University, Palo Alto, CA, (1988).

Charney, C., *Time to Market: Reducing Product Lead Time*, Society of Manufacturing Engineers, Dearborn, MI, (1991).

Dobson, G., U. Karmarkar, J. Rummel, "Batching to Minimize Flow Time on One Machine," *Management Science* **33**, (1987), 784-799.

Dobson, G., U. Karmarkar, J. Rummel, "Batching to Minimize Flow Times on Parallel Heterogeneous Machines," William E. Simon Graduate School of Business Administration, Working Paper No. 85-35, (1988).

Eppen G., and R. Martin, "Determining Safety Stock in the Presence of Stochastic Lead Time and Demand," *Management Science* **34** (1988), 1380-1390.

Hopp, W., M. Spearman, "Setting Safety Leadtimes for Purchased Components in Assembly Systems," Dept. Industrial Engineering and Management Sciences, Northwestern University, (1989), to appear in *IIE Transactions.*

Hopp, W., M. Spearman, D. Woodruff, "Practical Strategies for Lead Time Reduction," *Manufacturing Review* **3**, (1990), 78-84.

Karmarkar, U., "Lot Sizes, Lead Times, and In-Process Inventories," *Management Science* **33**, (1987), 409-418.

Karmarkar, U., "Manufacturing Lead Times, Order Release and Capacity Loading," working paper CMOM 98-04, Center for manufacturing and Operations Management, William E. Simon Graduate School of Business Administration, University of Rochester, (1989).

Karmarkar, U., S. Kekre, S. Kekre, S. Freeman, "Lotsizing and Lead Time Performance in a Manufacturing Cell," *Interfaces* **15**, (1985), 1-9.

Karmarkar, U., M. Lele, "The Manufacturing-Marketing Interface: Strategic and Operational Issues," working paper, William E. Simon Graduate School of Business Administration, (1989).

Kekre, S., and V. Udayabhanu, "Customer Priorities and Lead Times in Long Term Supply Contracts," *Journal of Manufacturing and Operations Management* **1**, (1988), 44-66.

Morton, T., and A. Vepsalainen, "Priority Rules and Leadtime Estimation for Job Shop Scheduling with Weighted Tardiness Costs," *Management Science* **33**, (1987), 1036-1047.

Ornek, M., and P. Collier, "The Determination of In-Process Inventory and Manufacturing Lead Time in Multi-Stage Production Systems," *International Journal of Operations and Production Management* **8**, (1988), 74-80.

Philiproom, P., L. Rees, B. Taylor, P. Huang, "Dynamically Adjusting the Number of Kanbans in a Just-in-Time Production System Using Estimated Values of Leadtime," *IIE Transactions* **19**, (1987), 199-207.

Ross, S., *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, (1983).

Schmenner, Roger W., "The Merit of Making Things Fast," *Sloan Management Review,* Fall Quarter, 11-17, (1988).

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

Shanthikumar, J., and U. Sumita, "Approximations for the Time Spent in a Dynamic Job Shop with Applications to Due Date Assignment," *International Journal of Production Research* **26**, (1988), 1329-1352.

Stalk, George and Thomas M. Hout, *Competing Against Time: How Time- Based Competition is Reshaping Global Markets,* The Free Press, New York, (1990).

Stidham, S., and R. Weber, "Monotonic and Insensitive Optimal Policies for Control of Queues with Undiscounted Costs," *Operations Research,* **37**, (1989), 611-625.

Thomas, Philip R., *Competitiveness Through Total Cycle Time: An Overview for CEO's*, McGraw-Hill, New York, (1990).

Thomas, Philip R., *Getting Competitive: Middle Managers and the Cycle Time Ethic*, McGraw-Hill, New York, (1991).

Vepsalainen, A., and T. Morton, "Improving Local Priority Rules with Global Lead Time Estimates," *Journal of Manufacturing and Operations Management* **1**, (1988), 102-118.

Wein, L.M., "Due-Date Setting and Priority Sequencing in a Multiclass M/G/1 Queue," *Management Science* **37**, (1991), 834-850.

Wijngaard, J., and S. Stidham, "Forward Recursion for Markov Decision Processes with Skip-Free-to-the-Right Transitions, Part I: Theory and Algorithms," *Mathematics of Operations Research,* **11**, (1986), 295-308.

Yano, C., "Setting Planned Lead Times in Serial Production Systems with Tardiness Costs," *Management Science* **33** (1987a), 96-106.

Yano, C., "Stochastic Leadtimes in Two-Level Assembly Systems," *IIE Transactions* **19**, (1987b), 371-378.