

# The value of sharing lead time information

GREGORY DOBSON and EDIEAL J. PINKER\*

*W. E. Simon School of Business, University of Rochester, Rochester, NY 14627-0100, USA*  
*E-mail: pinker@simon.rochester.edu*

Received December 2003 and accepted August 2005

---

The widespread adoption of Enterprise Resource Planning (ERP) systems has, among many other benefits, increased the ability of a firm to share operational data with customers. In this paper we analyze the factors that determine whether or not sharing a specific type of information, namely state-dependent lead time information, can benefit a firm. We develop a stochastic model of a custom-production environment, in which customers are handled on a first-come first-served basis but have differing tolerances for waiting. The firm has the option to share different amounts of information about the lead time a potential customer may incur. Although the information differs across scenarios, the reliability of that information in terms of the probability that a stated lead time is met is equal in the eyes of the customers. We derive conditions under which sharing more information with customers improves the firm's profits and the customers' experiences. We show that it is not always the case that sharing information improves the lot of the firm. We show that when customers' tolerances for waiting are more heterogeneous then the benefit to the firm from sharing lead time information increases. Our conclusion is that management should only authorize sharing detailed lead time information, be it through information system integration or frontline sales people, after a careful analysis of a customer's sensitivity to delay.

## 1. Introduction

Advances in information technology have greatly reduced the costs for a firm to share lead time information with customers. For example, it is now possible for a sales representative to instantly check the availability of a part in an inventory while responding to a customer call. In a custom-production environment a sales representative can also provide the customer with accurate lead time information based upon the actual number of orders already in the queue. In call centers it is possible to inform a customer of the anticipated queueing time. When a firm provides the customer with more information about the time it will take his/her request to be satisfied, it improves customer service. Whereas improving customer service should yield some benefits to the firm, in retaining customers and attracting new customers, the firm may also lose customers whose expectations are not met for particular service engagements. Therefore, it is not clear *a priori*, whether the firm or the customer reaps the benefits of sharing more information and thus how much information a firm should be willing to share.

In this paper, we focus on firms that provide a service to customers and seek to maximize the throughput of satisfied customer requests. To improve our understanding of how the sharing of lead time information with customers

affects a firm and its customers, we model the operation of a firm with an information system that provides customers with estimates of the lead time to satisfy a new request, based upon the number of requests already in the queue. We compare this firm to one that provides a single lead time quote to all customers without knowledge of the number of requests in the queue but based on long-term system performance. Both situations provide truthful quotes based on the available information. We compare these two levels of information sharing in terms of the throughput achieved by the firm and the service provided to the customer.

A naïve analysis would say that sharing information will lead to either a decrease or an increase in order submissions. If orders decrease then a traditional queueing view tells us that the throughput decreases and therefore the expected waiting time experienced by customers must also decrease. If orders increase then we expect that both the throughput and the expected waiting time experienced by customers will increase. In this paper we show that it may be possible to use the queueing information to not only increase the throughput but also decrease the expected waiting time for customers. Our results show that whether or not the firm benefits from sharing state-dependent lead time information will depend on the shape of the demand curve, i.e., the nature of a customer's sensitivity to waiting. We also show that sharing lead time information can cause the waiting time to go down, so that customers also benefit from information sharing. In this paper we model a single firm in isolation. Clearly, a firm's desire to share information would

---

\*Corresponding author

be affected by the information sharing policies of its competitors but analysis of the competitive problem is beyond the scope of this paper.

In Section 2 we review the relevant literature. In Section 3 we formulate our model. In Section 4 we derive conditions on the customer demand curve that guarantee that the firm benefits from sharing state-dependent lead time information. We also analyze the effect of sharing information on customers. In Section 5 we interpret the conditions on the demand curve derived in Section 4. In Section 6 we compare our results with the results of similar models in the literature. We conclude in Section 7 and highlight some areas for future research.

## 2. Literature review

In both manufacturing and services it is common for firms to compete intensely on the basis of price and lead time. When customers themselves are competing on lead time, for their own customers, they become increasingly sensitive not only to lead times but also to the reliability of quoted lead times. This effect is intensified in environments in which a customer is purchasing inputs for a lean-manufacturing operation. As a result there has been considerable research on how lead time quotes in manufacturing and services can be optimized to maximize a firm's performance.

Because queueing theory is typically used to model the lead times experienced by customers this problem has been modeled using queues in which different strategies for determining lead time quotes are compared or analyzed. One major modeling element that divides the literature is whether or not the lead time quotes are modeled as decision variables. There is an extensive literature on how to set and calculate lead times (or due dates) in production that views lead time quotes as a control mechanism. There is a smaller body of literature in which lead time quotes are viewed as information that is shared with the customer and the question of how much information to share is raised. Our work falls in this latter category.

Seidmann and Smith (1981), Wein (1991), Duenyas (1995), Duenyas and Hopp (1995), Palaka *et al.* (1998), Spearman and Zhang (1999), and also Kaupscinski and Tayur (2000) are all examples of work in which the firm has full control over the lead time information it makes available to a customer and deviations from the quotes result in some penalty. Papers in this area are distinguished by whether or not they model pricing, scheduling, and/or customer behavior. With increased integration of information systems within a supply chain, and widespread use of ERP systems, there are many situations in which suppliers have intentionally or inadvertently given up this control. Furthermore, in a competitive environment it is reasonable to expect that the firm will offer lead time quotes that satisfy some market standard for reliability, where, in this paper, reliability is defined as the probability of the actual lead time being

less than or equal to the quoted lead time. This definition of reliability is similar to the notion of a "service level" that commonly appears in the operations literature.

There is also a literature in which lead time information is shared rather than controlled. Whitt (1999) models two systems, one in which no lead time information is provided to customers and one in which exact state-dependent lead times are quoted. In Whang (1988, Ch. 5), a scenario in which state-dependent expected lead times are provided, is compared with providing long-run average lead time information. In both of these papers the comparison of information sharing scenarios is made for lead time quotes that have different probabilities of being satisfied.

In this paper, we model lead time quotes in such a way that the reliability of the quote, i.e., the fraction of time that the actual lead time is less than the quote, is held constant for all information scenarios compared. As is argued in Spearman and Zhang (1999), this measure of lead time quotes ("serviceability of jobs" in their language) is the measure in which practitioners are most interested. Ho and Zheng (2004) also model the information used by a customer when deciding to place an order with a firm as a lead time quote and probability of meeting that lead time. Kumar *et al.* (1997) investigate how customer satisfaction is impacted by a guarantee on the lead time. Their work also supports our model of providing a lead time estimate, which will be a (high) fractile of the lead time distribution. Finally, this approach captures situations in which suppliers have made lead time information directly accessible to their customers or environments in which the market imposes service-level constraints.

This constant-reliability model also makes it possible to generate comparisons of system performance across a spectrum of information sharing scenarios. Plambeck (2000) has considered a situation in which the firm quotes lead times that are 100% correct, a situation that she acknowledges is impossible, but can be approximated through a process termed "asymptotic compliance". In our model probabilistic compliance to lead time quotes is explicitly and precisely modeled. Similar to Whitt (1999), yet unlike most of the related literature, we also analyze the experience of the customer in terms of expected lead times and the variance of the lead time. In theory one could imagine situations in which a firm quotes an entire lead time distribution rather than a single number, but we have not observed this situation being addressed in the academic or practitioner literature.

## 3. The model

In this section we develop our model. The notation used in the paper is as follows.

### *Basic exogenous parameters*

- $\lambda$  = arrival rate of customers per unit time;
- $\mu$  = processing (service) rate;

- $\tau$  = lead time quote reliability parameter, between 0 and 1;
- $\alpha(l)$  = fraction of customers satisfied by the lead time quote  $l$ .

**Information related to scenario  $k$  and state  $i$**

- $S_k$  = scenario in which the firm provides state-dependent lead time quotes for states 0 through  $k - 1$  and an average lead time quote conditioned on being in state  $k$  or higher;
- $l_i$  = state-dependent lead time quote;
- $\bar{l}_k$  = lead time quoted to customers who arrive in states  $k$  and higher in scenario  $S_k$ ;
- $\lambda_i$  =  $\lambda\alpha(l_i)$  and is the order submission rate in scenarios  $S_k$ , state  $i < k$ , with lead time quote  $l_i$ ;
- $\bar{\lambda}_k$  =  $\lambda\alpha(\bar{l}_k)$  and is the order submission rate in scenarios  $S_k$ , states  $i \geq k$  with lead time quote  $\bar{l}_k$ ;
- $L_i$  = random system time for a customer in scenario  $S_k$  who submits an order when there are  $i < k$  orders in the system;
- $\bar{L}_k$  = random system time for a customer in scenario  $S_k$  who submits an order when there are  $k$  or more orders in the system;
- $A_k(l)$  = supply curve, i.e., the order submission rate for which the firm would offer a lead time quote  $l$  in state  $k$ ;

**Performance measures for the system under scenario  $S_k$**

- $R_k$  = throughput;
- $W_k$  = expected waiting time;
- $V_k$  = variance in waiting time.

A stream of customers arrives according to a Poisson process, with an exogenously specified rate  $\lambda$  and with each customer having a unique request. Customers then decide whether or not to submit their requests, based on the quoted lead time,  $l$ , which may depend on the state of the manufacturing system and thus vary over time, and the quote's reliability (or serviceability)  $\tau \in [0, 1]$  (see Fig. 1). That is, the firm is quoting a lead time  $l$ , that will be met  $100\tau\%$  of the time. A customer,  $c$ , derives a utility from a  $(l, \tau)$  pair,  $U_c(l, \tau)$ . If that utility exceeds some threshold,  $u_c$ , namely  $U_c(l, \tau) \geq u_c$ , then he/she will submit a request. This model of the customer decision process is similar to that of Ho and Zheng (2004). Customers who decide not to submit requests, go elsewhere, and hold no ill will against the firm. The function  $U_c(l, \tau)$  is decreasing in  $l$  for all customers and for all  $\tau$  so a customer's decision of whether or not to place an order,  $I(\{U_c(l, \tau) \geq u_c\})$ , i.e., the indicator function of the event, is also decreasing in  $l$ .

Customers are heterogeneous in their willingness to wait, and so we define  $\alpha(l, \tau) \in [0, 1]$  as the fraction of customers who decide to place an order when presented with  $(l, \tau)$ . From the discussion above, the function  $\alpha(l, \tau)$  is decreasing. We further assume that it is continuous. In this paper,

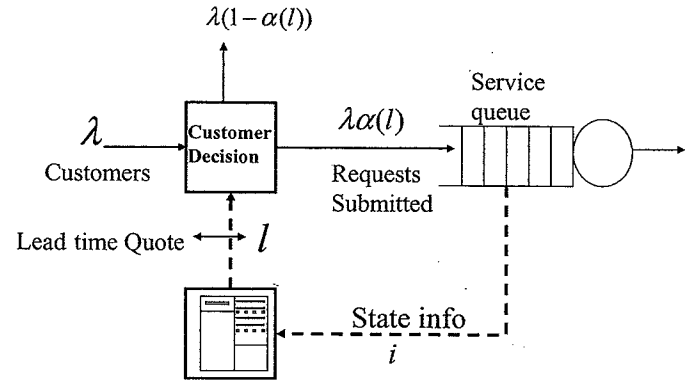


Fig. 1. A schematic of the customer and information flows.

we assume that  $\tau$  is an exogenous market standard and thus we suppress it. When the firm is quoting a lead time  $l$  the customer submission rate is thus  $\lambda\alpha(l)$  and we call  $\lambda\alpha(l)$ , the demand curve.

We model the firm's production system as an M/M/1 queue (an assumption we relax later in the paper) with a first-come first-served service discipline and possibly state-dependent arrival rates. We assume that the firm wants to maximize the throughput of satisfied requests. Customers understand that the firm cannot predict the actual waiting time with certainty and have a common exogenous parameter,  $\tau \in [0, 1]$ , which is the lead time quote reliability they require. The firm must state a lead time that is achieved  $100\tau\%$  of the time. What this means depends on the information that the firm is sharing. We model two scenarios:  $S_0$  (no state information) and  $S_\infty$  (state-dependent information for all states).

**3.1. Scenario  $S_0$**

In this scenario the firm provides all the customers with a single lead time,  $\bar{l}_0$ , all the time, based only on the long-run performance, and  $\bar{l}_0$  is the minimum lead time that can be met ( $100\tau\%$ ) of the time for a randomly arriving customer. Because all customers are quoted the same lead time  $\bar{l}_0$  there is a single request submission rate  $\bar{\lambda}_0 = \lambda\alpha(\bar{l}_0)$ , and thus the system operates as a simple M/M/1 queue with service rate  $\mu$  and arrival rate  $\bar{\lambda}_0$ . The lead time experienced by a customer in scenario  $S_0$  is  $\bar{L}_0$  and its CDF is given by  $G_{\bar{L}_0}(l) = 1 - e^{-\mu - \bar{\lambda}_0 l}$  (Kleinrock, 1975). The lead time quote  $\bar{l}_0$  is selected so that:

$$\Pr\{\bar{L}_0 \leq \bar{l}_0\} = G_{\bar{L}_0}(\bar{l}_0) = 1 - e^{-(\mu - \lambda\alpha(\bar{l}_0))\bar{l}_0} = \tau. \quad (1)$$

**3.2. Scenario  $S_\infty$**

In this scenario the firm provides the customer with a state-dependent lead time quote  $l_i$  based upon the number of requests,  $i$ , already in the system and the  $(100\tau)\%$

confidence level used in scenario  $S_0$ . The *state-dependent* order submission rate is  $\lambda_i = \lambda\alpha(l_i)$ . The state-dependent lead time quotes,  $l_i$ , are determined by the 100 $\tau$  percentile of a random variable  $L_i$ , with an  $(i + 1)$ th-order Erlang distribution (an assumption relaxed later in the paper), namely:

$$\Pr\{L_i \leq l_i\} = \int_0^{l_i} \frac{\mu^{i+1} x^i e^{-\mu x}}{i!} dx = \tau. \quad (2)$$

Thus, the system can be characterized as a birth-death process in which the steady-state probability of being in a state with  $i$  customers in the system is  $\pi_i$  where  $\lambda_i \pi_i = \mu \pi_{i+1}$ . Note that  $\lambda$  may be greater than  $\mu$  because not all arrivals submit requests and join the queue. Depending upon the distribution of the customers' lead time requirements, the submission rate may also be significantly less than  $\lambda$ . Also note from Equations (1) and (2) that the lead time quotes, and therefore submission rates and throughput, in both scenarios, are dependent on  $\tau$  and  $\alpha(l)$ , which we assume are exogenously determined. This model of a firm is similar to the deadline delay cost that appears in Dewan and Mendelson (1990) and would be particularly applicable when the customer is operating in a just-in-time environment.

#### 4. Analysis

Our goal is to compare the two information scenarios described above, from the perspectives of the firm and the customers. The performance metric for the firm is throughput whereas customers are interested in service-oriented performance metrics. We are interested in both the means and the variances of a customer's lead time because changes in either may have important operational effects on a customer's supply chain. In what follows we investigate how the information scenario affects all of these performance measures. To highlight the need for our proposed approach in Fig. 2 we first consider an illustrative example where  $\mu = 1$  customer/hour,  $\lambda = 0.8$  customers/hour,  $\tau = 0.95$ , and  $\alpha(l) = 1 - l/10$ .

In Fig. 2, the horizontal axis shows the current state of the system. The two solid lines depict the lead time quotes for scenarios  $S_0$  and  $S_\infty$ , and are plotted relative to the left-hand vertical axis. Note that the lead times for the  $S_\infty$  scenario increase with the number of customers in the system, whereas the  $S_0$  lead time is constant. When there are very few customers in the system, the state-dependent lead time quotes are shorter than the constant lead time quotes. However, as the system becomes more crowded the  $S_\infty$  lead time quotes exceed the lead time quotes of  $S_0$  influencing the number of customers served.

The two dashed lines depict the order submission rate under scenarios  $S_0$  and  $S_\infty$  and are plotted relative to the right-hand vertical axis. We can see that when we

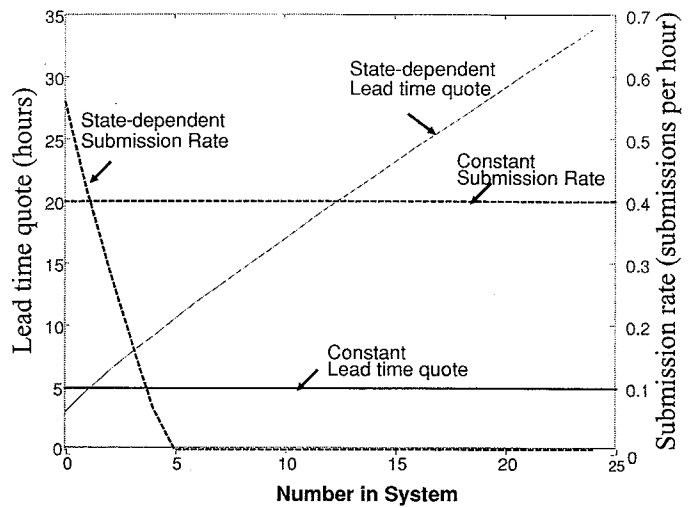


Fig. 2. Constant and state-dependent arrival rates and lead time quotes.

are in states where the  $S_\infty$  quotes are lower than the  $S_0$  quote,  $S_\infty$  has more customers entering the system than  $S_0$ , meaning a greater throughput. Determining whether or not  $S_\infty$  is superior to  $S_0$  requires determining the circumstances under which the  $S_\infty$  system spends sufficient time in states to the left of the crossing point of the lead time curves.

#### 4.1. Definition of intermediate scenarios, $S_k$

As a mechanism for deriving our results we introduce intermediate information scenarios,  $S_k$ : in states 0 through  $k - 1$  the firm gives the customers state-dependent lead time quotes  $l_i$ , which result in order submission rates  $\lambda\alpha(l_i)$ , whereas in states  $k$  and higher the firm gives the customers a single average lead time quote,  $\bar{l}_k$ , which results in an order submission rate at  $\lambda\alpha(\bar{l}_k)$ . For example in scenario  $S_1$  the firm either informs a newly arrived customer that the production system is empty and quotes a lead time based on the service time for the customer, or if the system is not empty, the firm quotes a lead time based on the average state, conditioned on there being at least one order in the system, using the reliability requirement  $\tau$ . Clearly in scenario  $S_1$  more information is provided than in scenario  $S_0$ , with  $S_k$  providing more information than  $S_{k-1}$ .

The quote  $l_i$  is determined as in scenario  $S_\infty$  (see Equation (2)). The quote  $\bar{l}_k$ , for states  $k$  and higher is selected so that  $\Pr\{\bar{L}_k \leq \bar{l}_k\} = \tau$ , where  $\bar{L}_k$  is the actual lead time experienced by a request that is placed when the system is in state  $k$  or higher with an order submission rate of  $\lambda\alpha(\bar{l}_k)$  and a service rate of  $\mu$ . Note that the probability distribution of  $\bar{L}_k$  is the probability distribution of a  $k$ th-order Erlang convolved with the probability distribution function of the system time in an M/M/1 queue with service rate  $\mu$

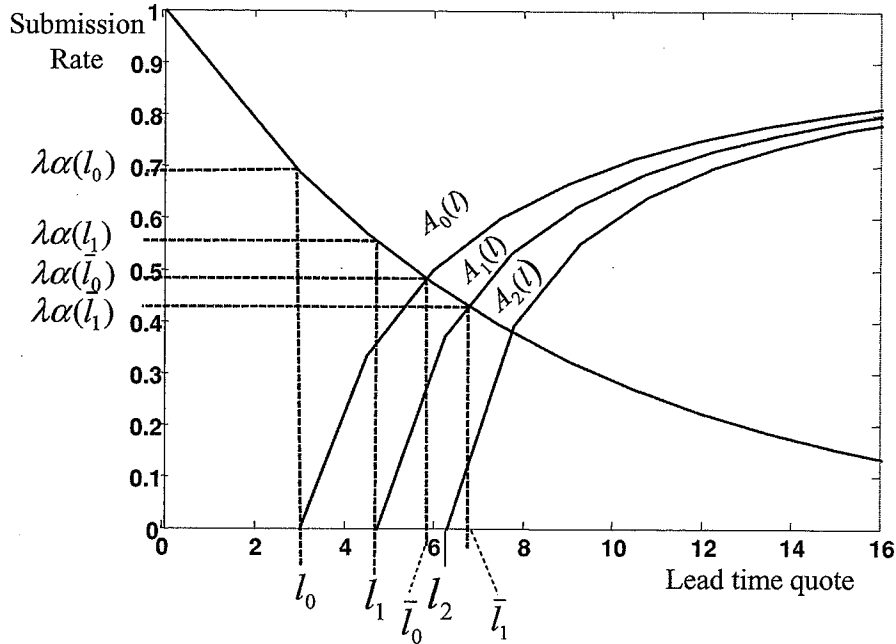


Fig. 3. The “demand” and “supply” curves that define  $\bar{l}_k$ .

and arrival rate  $\lambda\alpha(\bar{l}_k)$ . Therefore,  $\bar{l}_k$  is the solution to:

$$\Pr\{\bar{L}_k \leq \bar{l}_k\} = \int_0^{\bar{l}_k} \frac{\mu^k x^{k-1} e^{-\mu x}}{(k-1)!} (1 - e^{-\mu - \lambda\alpha(\bar{l}_k)(\bar{l}_k - x)}) dx = \tau. \quad (3)$$

We depict the determination of  $\bar{l}_i$  and  $l_i$  graphically in Fig. 3 for the case of  $\mu = 1$  and  $\tau = 0.95$ . The upward sloping curves represent order submission rates that would have caused the firm to generate the associated lead time quote when at least zero, one, and two customers are already in the system. In other words, given  $\mu$  and  $\tau$  these curves are generated by plotting the pairs  $(l, y)$  that satisfy the following equation, for different values of  $k$ :

$$\Pr\{\bar{L}_k \leq l\} = \int_0^l \frac{\mu^k x^{k-1} e^{-\mu x}}{(k-1)!} (1 - e^{-(\mu - y)(l - x)}) dx = \tau. \quad (4)$$

We define the function  $A_k(l)$  as the  $y$  satisfying Equation (4). These upward sloping curves,  $A_k(l)$ , can be thought of as “supply curves”. The downward sloping solid curve is the “demand curve”,  $\lambda\alpha(l)$ . Recall that  $l_i$  is the lead time quote given at a customer arriving at a system in state  $i$ . The lead time quote at which  $\lambda\alpha(\bar{l}_i) = A_i(\bar{l}_i)$  is  $\bar{l}_i$ . For example, in Fig. 3,  $\bar{l}_0$  is the lead time quote that solves Equation (3) for scenario  $S_0$ , i.e., for  $k = 0$  and  $\lambda\alpha(\bar{l}_0)$  is the order submission rate for that lead time quote.

**Lemma 1.** *The supply curves,  $A_k(l)$ , have the following properties:*

- (a).  $A_k(l) \leq \mu$ ;
- (b).  $\frac{d}{dl} A_k(l) \geq 0$ ;
- (c).  $A_{k-1}(l) \geq A_k(l)$ .

**Proof.** See Appendix A. ■

In the following lemma we state some useful relationships among the order submission rates generated by different lead time quotes.

**Lemma 2.**

$$\begin{aligned} l_i &\leq \bar{l}_i \leq \bar{l}_{i+1} && \forall i \geq 0 \\ \lambda\alpha(l_i) &\geq \lambda\alpha(\bar{l}_i) \geq \lambda\alpha(\bar{l}_{i+1}) && \forall i \geq 0. \end{aligned}$$

**Proof.** Follows directly from definitions. ■

In the following, we analyze and compare the system performance under scenarios  $S_0$  and  $S_\infty$  from the perspective of the firm and the customers.

#### 4.2. Firm throughput

We denote the throughput of the firm in scenario  $S_k$  as  $R_k$ . We now define a set of conditions  $C_k$  that play an important role in our main results. Define  $C_k$  as the condition that:

$$\frac{\alpha(l_k)}{\alpha(\bar{l}_k)} \geq \frac{1 - \alpha(\bar{l}_{k+1})(\lambda/\mu)}{1 - \alpha(\bar{l}_k)(\lambda/\mu)}. \quad (5)$$

Furthermore, define  $C$  as the condition that  $C_k$  holds for all  $k$ . This set of conditions will be further discussed in our proof of Proposition 1 and we simply note at this point that it is related to the shape of the demand curve. In Section 5 we develop an interpretation of these conditions.

**Proposition 1.** *For a given  $k$ ,  $R_k \leq R_{k+1}$  if and only if  $C_k$  holds and if  $C$  holds then,  $R_0 \leq R_\infty$ .*

**Proof.** See Appendix A. ■

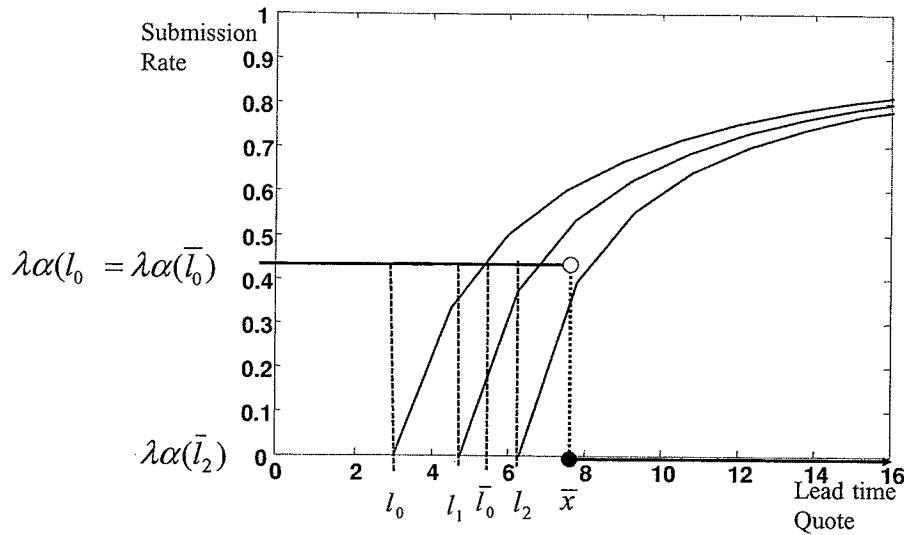


Fig. 4. A step function demand curve.

Condition *C* is therefore a sufficient condition for throughput to be higher when state-dependent lead time quotes are supplied to the customer. We have found numerically many examples, including exponential, of demand curves for which *C* holds and  $R_k$  is increasing in  $k$  and so sharing information increases throughput. Some examples of these appear in Section 5. However, using Proposition 1 as a guide, we can construct interesting examples of situations in which sharing information reduces throughput.

In Fig. 4 we depict a demand curve that is a simple step function that has a positive value for all lead times less than  $\bar{x}$  and is zero otherwise. This demand function represents a

situation in which all customers are indifferent about lead times below a certain delay threshold,  $\bar{x}$ . In this example  $\alpha(l_0) = \alpha(\bar{l}_0) = \alpha(l_1) = \alpha(\bar{l}_1) = \alpha(l_2)$  while  $\alpha(\bar{l}_2) = 0$ . The implication is that in scenarios  $S_0$  and  $S_1$  the throughput is equal, namely  $R_0 = R_1$ , whereas adding another state of information, i.e., scenario  $S_2$  reduces throughput,  $R_2 < R_1$ .

We can also construct a demand curve such that  $\alpha(\bar{l}_k) = \alpha(l_k)$  (see Fig. 5), where every additional state, for which the firm provides information, reduces the throughput. We expect diminishing benefits, but in this case additional information sharing always has a negative value. In Fig. 6 we plot  $R_k$  against  $k$ . The two curves represent the expected

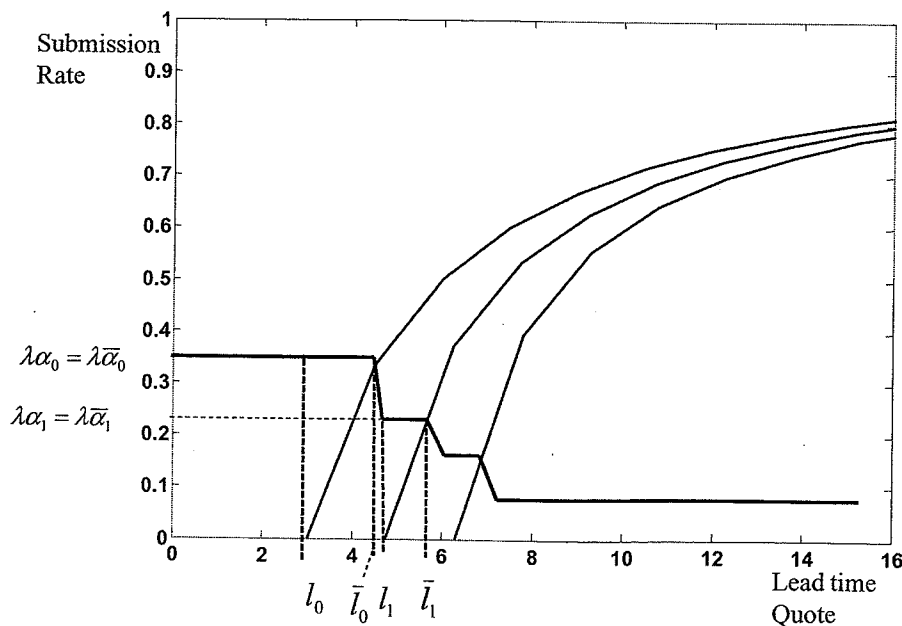


Fig. 5. A demand curve which results in information sharing having a negative value for the firm.

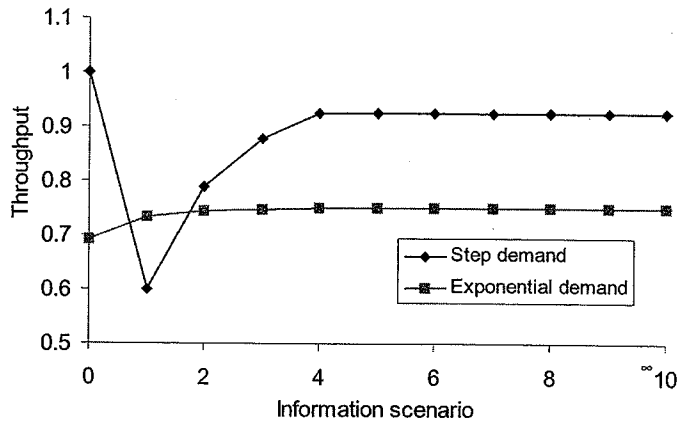


Fig. 6. Throughput as a function of information scenario for two demand curves.

throughput for a step function and a negative exponential demand curve. See Appendix B for details on the parameter values used to generate Figs. 6 and 7. The step function is an example of a demand curve for which the throughput is higher in scenario  $S_0$  than in scenarios where state information is shared.

### 4.3. Expected value and variance of customer wait

State-dependent lead time information causes the order submission rate to be higher in states with shorter lead times than in states with longer lead times. Intuitively, we anticipate that the expected lead time,  $W_k$ , experienced by a random customer submitting an order will be shorter when information is shared than when only average lead time information is given. Similarly, because order submissions are concentrated in the lower-numbered (and thus less waiting time) states we also expect the variance of lead time,  $V_k$ , to be smaller with information sharing than without. For scenarios  $S_0$  and  $S_1$  we have the following result.

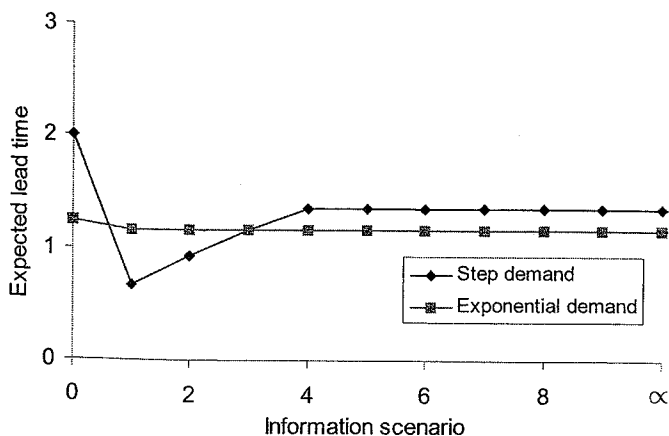


Fig. 7. Expected lead time as a function of information scenario.

**Proposition 2.**  $W_0 \geq W_1$  and  $V_0 \geq V_1$ . In particular  $W_1$  is given by  $1/(\mu - \lambda_1)$ .

**Proof.** See Appendix A. ■

According to Proposition 2, the expected lead time in scenario  $S_1$  is independent of the state 0 order submission rate  $\lambda_0$ . Taken together with Proposition 1 we see that even if the throughput is higher in scenario  $S_1$  than  $S_0$  it is possible for the expected lead time to be lower. What happens in scenarios  $S_k (k > 1)$ ?

In Fig. 7 we plot the expected lead time as a function of information scenario for a step function and a negative exponential demand curve. For both demand curves the expected lead time is lower when some state-dependent information is shared than when none is shared ( $S_0$ ), and for  $k > 0$  lead times can increase with additional state information.

The expected lead time is lower with state information because a higher proportion of customers who place orders place them when the system is less congested, e.g., when it's empty. The expected lead time can increase with additional state information beyond scenario  $S_1$  because additional state information increases order submission rates in the additional states and increases throughput. Comparing Figs. 6 and 7 for the exponential demand curve we see that the expected lead time is lower in scenario  $S_k (k > 0)$  than in  $S_0$ , i.e.,  $W_k < W_0$ , even though the throughput is greater, i.e.,  $R_k > R_0 (k > 0)$ . The results for lead time variance are very similar to those for the expected lead time.

From the examples described above, we can see that, given the appropriate demand curve, it is possible to increase a firm's throughput while simultaneously decreasing a customer's expected lead time and variance of lead time, by sharing state-dependent lead time information.

## 5. Interpretation of conditions $C_k$

In this section we relate the conditions  $C_k$  to the heterogeneity across the customers' lead time requirements and then to the value of information sharing. Define demand curves:

$$D_t(l) = \begin{cases} \lambda^t(1 - (l/10)^{t+1}) & \text{for } t \geq 0 \text{ and } l \leq 10, \\ \lambda^t(1 - l/10)^{1-t} & \text{for } t < 0 \text{ and } l \leq 10, \\ 0 & \text{otherwise.} \end{cases}$$

where  $\lambda^t$  will be selected so that the throughput in scenario  $S_0$ , i.e.,  $R_0$ , is constant, which will make the comparisons we do later sensible. (Through numerical experiments we have found that for a given demand curve  $\lambda\alpha(x)$ ,  $R_\infty/R_0$  is a decreasing function of  $\lambda$ . Therefore, because changing the shape of the demand curve also changes the overall arrival rate to the system it is necessary to normalize the arrival rate to isolate the effect of demand curve shape on  $R_\infty/R_0$ .)

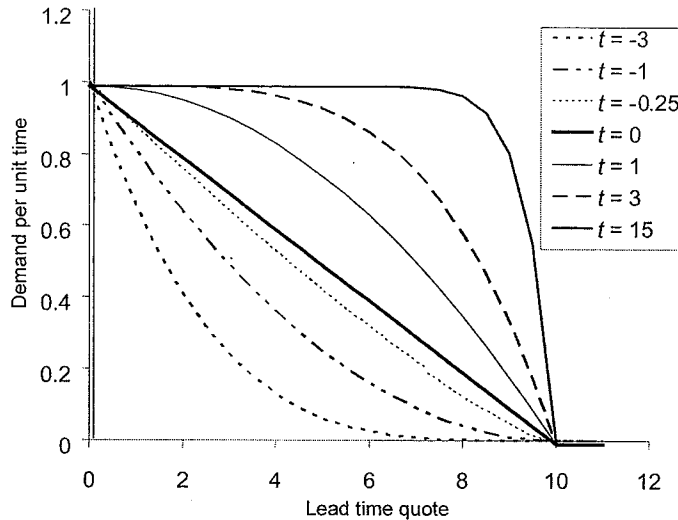


Fig. 8. A family of demand curves.

Some members of this family of demand curves are plotted in Fig. 8, with  $\lambda^t = 1$  for all  $t$ . We can see in this figure that for  $t = 15$  the customers are homogeneous for lead time quotes 1–4, whereas for  $t < 15$  the customers are heterogeneous over these same lead time quotes, and as  $t$  decreases they are increasingly heterogeneous over the lower lead time quotes, quotes 1–4.

For each  $D_t(l)$ , we can calculate  $\gamma$ , the fraction of time the system is in a state for which the condition  $C_k$  is satisfied, namely  $\gamma = \sum_k \pi_k I_k$  where  $I_k$  is an indicator variable with:

$$I_k = \begin{cases} 1 & \text{if } C_k \text{ is true,} \\ 0 & \text{if } C_k \text{ is false.} \end{cases}$$

Recall that  $C$  is the condition that  $C_k$  holds for all  $k$ ; this means that  $\gamma$  can be used as a measure of the conformity of the demand curve to condition  $C$ . When  $\gamma = 1$ , condition  $C$  holds. When  $\gamma < 1$ , condition  $C$  is violated for some states. In Fig. 9 we see that as  $t$  increases the system is in a state in which  $C_k$  is violated more often, whereas in Fig. 8 we see that as  $t$  increases the customers are more homogeneous in their sensitivity to the lower lead time quotes. Thus, we see a connection between the homogeneity in customers at lower lead time quotes and the fraction of time that condition  $C$  is violated.

Next we turn to the impact of demand curve shape on throughput. In Fig. 10 we plot  $R_\infty/R_0$  as a function of  $t$  for two different values of  $R_0$ , 0.4 and 0.6. The ratio reveals the value gained by the firm for sharing additional information. We can see that for both values of  $R_0$  the value of state information decreases in  $t$ . Looking at Figs. 9 and 10 together, we see that the increased violation of the condition  $C_k$  corresponds to the poorer performances of scenario  $S_\infty$  relative to scenario  $S_0$ , that is we see a direct relationship between the fraction of time conditions  $C_k$  are met and the relative value of more information sharing. This means that, for the family of demand curves we consider, the more

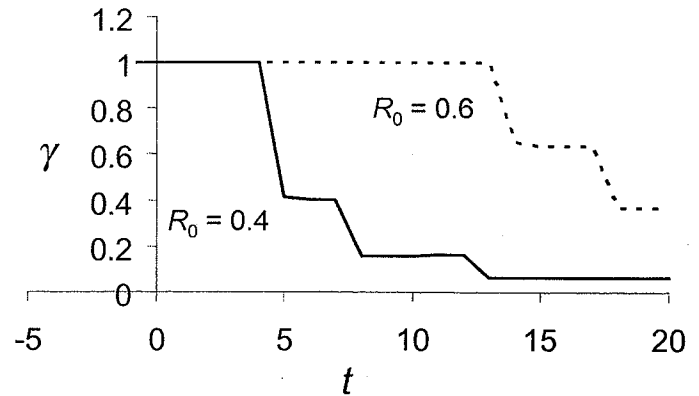


Fig. 9.  $\gamma$  versus  $t$ .

the demand curve represents homogeneous customers (for short lead time quotes) the less it satisfies conditions  $C_k$  and the less value the firm accrues from information sharing. (Note that we can define an alternative measure of the conformity of the demand curve to  $C$ ,  $EC$ , where  $EC$  is the expected value of the left-hand side of condition  $C_k$  minus the right-hand side. The lower the value of  $EC$  the greater the degree to which condition  $C$  is violated. We found that  $EC$  behaved in a similar manner to  $\gamma$  as  $t$  was varied.)

Incidentally, as  $t$  increases  $D_t(l)$  looks more like a step function. Thus, the step function depicted in Fig. 4 is not a pathological case but an extreme form of homogeneity across customers, and there is a direct connection between homogeneity/heterogeneity and the value derived by the firm from lead time information sharing. Also, in Fig. 10 we see that the arrival rate to the system plays a role. When throughput is lower ( $R_0$  fixed at 0.4) scenario  $S_0$  performs better relative to  $S_\infty$  than when throughput is higher ( $R_0$  fixed at 0.6).

What is the intuition behind the role of customer heterogeneity? Let us compare the no information  $S_0$  case to the

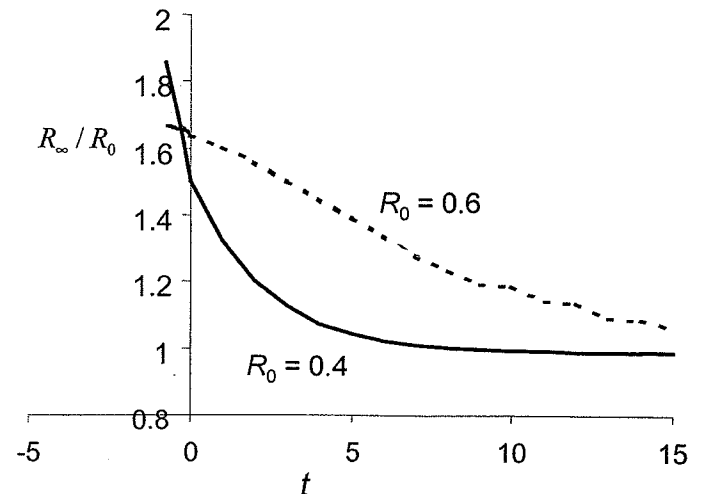


Fig. 10.  $R_\infty/R_0$  versus  $t$ .

$S_\infty$  case for the two demand curves in Fig. 8 when  $t = 3$  and  $t = -3$ . For  $t = 3$ , consider changing from  $S_0$  to  $S_\infty$ . For the lower number states the firm offers a lower lead time in  $S_\infty$  than the average in  $S_0$ , and has a chance to gain more customers. Because there is little difference between customers little is gained. For the higher numbered states the firm offers worse lead times compared to  $S_0$ , and loses many customers because this is where the demand curve falls off. Alternatively, for  $t = -3$ , in the lower-numbered states, where the firm can offer better lead time estimates in  $S_\infty$  than in  $S_0$ , there is significant heterogeneity among customers, which allows the firm to capture new customers. In the higher-numbered states, where the firm can offer only longer lead times than the  $S_0$  case, customers are fairly homogeneous so sharing information does not significantly affect the profit.

To verify the robustness of our results outside the M/M/1 framework we experimented with service time distributions drawn from the gamma family. We selected the parameters of the gamma pdf so that the Coefficient of Variation (CV) of the service time was 0.5 and 1.5 while holding the mean service time constant at 1 as is the case in the previous experiments. Since an analytical expression for the system time distribution of M/G/1 queues does not exist we have to rely on the Laplace transform of the system time distribution, which is known (Kleinrock, 1975, p. 407), to calculate the correct lead time quotes for scenarios  $S_0$  and  $S_\infty$ . The lead time quotes that satisfied the  $\tau\%$  requirement were found using a numerical inversion procedure for Laplace transforms outlined in Abate and Whitt (1995). The throughput in scenario  $S_\infty$ ,  $R_\infty$ , must then be determined using Monte Carlo simulation. Since the M/G/1 queue is not Markovian the expression for the conditions  $C_k$  does not hold anymore.

In Fig. 11 we plot  $R_\infty/R_0$  as a function of  $t$  (the demand curve shape parameter)  $R_0 = 0.4$  and three different service

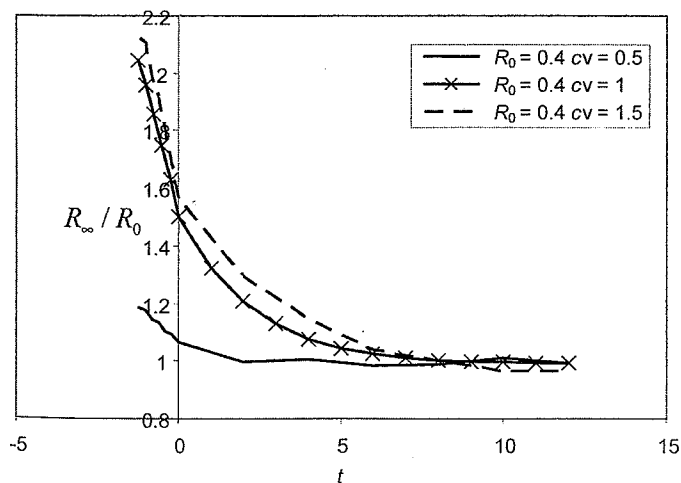


Fig. 11.  $R_\infty/R_0$  as a function of demand curve shape,  $t$ , for gamma service time distributions with CV values of 0.5, 1.0, and 1.5 and a mean of 1.

time distributions. We can see that the results are quite similar. As the shape of the demand curve becomes more like a step function reflecting a more homogenous customer base the benefit of sharing information decreases.

## 6. Discussion

In this section we give a detailed comparison with previous work in this area and show that while some of our results are consistent with those in the literature, others are different. One of our results is the identification of some situations in which increased information sharing may in fact reduce the firm's throughput. Whang (1988) has also shown an example in which providing state-dependent average lead time information generates a lower value to the system than providing only average lead times. Whereas his model is one of total system performance, combining customer and firm utilities, ours maximizes throughput the examples have a similar flavor in that the objective is worse with state-dependent information sharing. However, in his example throughput is indeed higher with state-dependent information whereas in ours throughput can be lower. His counter-example to the intuition that "more information is better" seems to be a result of his choice of customer value functions. Our Proposition 1 gives general conditions for when sharing increasing amounts of information reduces throughput.

Furthermore, Whang compares two information scenarios which are not really comparable from a customer's perspective. In particular, for the case where he provides only average lead times the probability of being on time is  $1 - 1/e$  whereas for the case where he provides state-dependent average lead times the probability of being on-time is bounded from above by  $1 - 1/e$  and converges to 0.5.

Whitt (1999) develops a model similar to ours, of an M/M/s queuing system with limited waiting space applicable to call center management. In his model arriving customers have "waiting time tolerances" that are chosen from an exponential distribution forming the single demand curve he considers. He compares two information scenarios. In the first scenario, which we denote WN (Whitt—No information), customers do not receive any information about the system delay. A customer who enters the system with no information will renege if his/her wait exceeds his/her waiting time tolerance. In the second scenario, WI (Whitt—Information), the customer receives exact information about the delay based upon the orders ahead of him/her in the queue, and he/she will balk (not submit his/her order) if the delay is longer than his/her tolerance. It is demonstrated that throughput is greater with less information than with more information.

This is an interesting result for our discussion because, if we order our scenarios  $S_0$  and  $S_\infty$  and Whitt's scenarios, WN and WI, from least information to greatest information we obtain the ordering  $WN < S_0 < S_\infty < WI$ . One

would think that the result that throughput decreases from WN to WI would be generalizable to throughput decreasing monotonically with increasing information, i.e.,  $R_{WN} \geq R_0 \geq R_\infty \geq R_{WI}$ . However, we saw in Fig. 7, the throughput in scenario  $S_\infty$  is higher than in scenario  $S_0$  with an exponential customer demand curve, i.e.,  $R_0 \leq R_\infty$ . On the other hand Proposition 1 demonstrates that  $R_0 < R_k, k > 0$  is not true for all demand curves.

Whitt's model includes a no information case with renegeing which is best suited to a call-center setting rather than the industrial setting we consider in our paper. Furthermore, we construct our information scenarios using common reliability levels that are more comparable. In Whitt, the full information scenario is heuristically modeled as being memoryless, which introduces additional distortions into the comparisons. The comparison of our results to Whitt's results suggests two lessons. First, the effect of information sharing is sensitive to modeling assumptions. Second, it is not possible to draw general conclusions about the effect of information sharing from only one demand function.

In Duenyas and Hopp (1995) the optimal lead time quote policy for the firm is shown to be state dependent and increasing in the state. From the perspective of the optimal control of a queue, in which there is a penalty for being late, this result is not surprising. This result appears to contradict our Proposition 1, which shows that increasing the number of states for which a unique lead time quote is given does not always improve throughput. However, we are modeling a situation in which the firm cannot choose the lead time quote to provide but must provide quotes that are achieved  $100\tau\%$  of the time. In Duenyas and Hopp (1995) the firm chooses the profit maximizing lead time quotes, i.e., they trade-off reliability penalties for throughput. The premise of this paper is that increased information system integration will make it more difficult for a firm to choose lead time quotes that do not meet a standard for reliability.

The distributions of lead times under the two information scenarios are different and so  $100\tau\%$  reliability in scenario  $S_0$  does not lead to exactly the same customer experience as a  $100\tau\%$  reliability in scenario  $S_\infty$ . However, these two information scenarios are certainly more comparable with lead time quotes based on the  $\tau$ -fractile than on the mean, for the reasons discussed above.

## 7. Conclusions

The problem of sharing state-dependent lead time information with customers has been modeled within a variety of contexts in the literature. We have developed a model that we believe, frames the problem in a way that is most pertinent to managers. The use of more information in decision making should lead to improved performance, however, it is not clear *a priori* who will receive this benefit, the firm or the customer. We have found that in many cases providing state-

dependent lead time information is better than information based upon long-run lead time distributions for both firm and the customer. It can increase a firm's throughput, while, in some cases, simultaneously reducing the customer's expected waiting time and the variance in the waiting time. On the other hand, we have also found that these benefits do not hold for general or even some reasonable customer demand curves and that, in light of previous work done on this subject, the possible benefits of more information are highly sensitive to modeling assumptions. In particular we have seen that a firm's throughput tends to increase with information sharing when customer lead time requirements are heterogeneous in lower-numbered states (ones where state-dependent lead time estimates would be less than the  $S_0$  scenario), and they tend to decrease when customer tolerances are homogeneous in these same states.

When a firm's information systems are integrated using ERP-type systems a byproduct is the ability to easily provide customers with detailed lead time information. This information may be transmitted to the customer via front-line sales people, software agents, or through extranets linking supply chain members. Our modeling and analysis gives evidence that this byproduct is not necessarily beneficial to the firm and must be carefully considered. A firm that does not have a clear understanding of its customers' preferences may in fact reduce throughput in the name of providing better service or supply chain integration.

Turning to future work, a characterization of the class of smooth demand curves that satisfy condition C would be a useful technical contribution. In this paper we have considered the capacity as being as fixed. The problem of determining the optimal capacity endogenously under information sharing is another area for future research. An additional extension would be to include pricing and competition between firms. Similar models could be developed of firms competing for customers in which the firms must choose an information strategy, capacity level, and prices.

## References

- Abate, J. and Whitt, W. (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing*, 7(1), 36-43.
- Dewan, S. and Mendelson, H. (1990) User delay costs and internal pricing for a service facility. *Management Science*, 36, 1502-1517.
- Duenyas, I. (1995) Single facility due date setting with multiple customer classes. *Management Science*, 41, 608-619.
- Duenyas, I. and Hopp, W. (1995) Quoting customer lead times. *Management Science*, 41, 43-57.
- Ho, T. and Zheng, Y.S. (2004) Setting customer expectation in service delivery: An integrated marketing-operations perspective. *Management Science*, 50, 479-488.
- Kaupscinski, R. and Tayur, S. (2000) Dynamic capacity reservation in a make-to-order environment. Working paper. GSIA, Carnegie Mellon University. Pittsburgh, PA 15213, USA.
- Kleinrock, L. (1975) *Queueing Systems Volume I: Theory*, Wiley, New York, NY.

Kumar, P., Kalwani, M.U. and Dada, M. (1997) The impact of waiting time guarantees on customers' waiting experiences. *Marketing Science*, **16**, 295–314.

Palaka, K., Erlenbacher, E. and Kropp, D. (1998) Lead time setting, capacity utilization, and pricing decisions under lead time dependent demand. *IIE Transactions*, **30**, 151–163.

Plambeck, E. (2000) Pricing, leadtime quotation and scheduling in a queue with heterogeneous customers. Working paper. Stanford University Graduate School of Business, Stanford, CA, 94305, USA.

Seidmann, A. and Smith, M. (1981) Due dates assignments for production systems. *Management Science*, **27**, 571–581.

Spearman, M. and Zhang, R. (1999) Optimal lead time policies. *Management Science*, **45**, 290–295.

Wein, L. (1991) Due date setting and priority sequencing in a multi-class M/G/1 queue. *Management Science*, **37**, 363–374.

Whang, S. (1988) Pricing computer services: incentive, information and queueing effects Ph.D. dissertation, University of Rochester, Rochester, NY, 14627, USA.

Whitt, W. (1999) Improving service by informing customers about anticipated delays. *Management Science*, **45**, 192–207.

## Appendices

### Appendix A

**Lemma 1.** *The supply curves,  $A_k(l)$ , have the following properties:*

- (a)  $A_k(l) \leq \mu$ ;
- (b)  $\frac{d}{dl} A_k(l) \geq 0$ ;
- (c)  $A_{k-1}(l) \geq A_k(l)$ .

#### Proof.

(a). Recall that  $A_k(l)$  is defined as the solution to:

$$\begin{aligned} \Pr\{\bar{L}_k \leq l\} \\ = \int_0^l \frac{\mu^k x^{k-1} e^{-\mu x}}{(k-1)!} (1 - e^{-(\mu - A_k(l))(l-x)}) dx = \tau \end{aligned} \quad (\text{A1})$$

If  $A_k(l) > \mu$  then we would get negative probabilities in Equation (A1).

(b). Given  $\tau$ , we want to compare  $A_k(l)$  and  $A_k(l + \varepsilon)$ . First, to simplify notation we define:

$$B_k(x) = \frac{\mu^k x^{k-1} e^{-\mu x}}{(k-1)!}.$$

For a fixed  $\tau$ , we have from Equation (A1):

$$\begin{aligned} & \int_0^l B_k(x) (1 - e^{-(\mu - A_k(l))(l-x)}) dx \\ &= \int_0^l B_k(x) (1 - e^{-(\mu - A_k(l+\varepsilon))(l+\varepsilon-x)}) dx \\ & \quad + \int_l^{l+\varepsilon} B_k(x) (1 - e^{-(\mu - A_k(l+\varepsilon))(l+\varepsilon-x)}) dx. \end{aligned} \quad (\text{A2})$$

From part (a) above we have that  $(1 - e^{-(\mu - A_k(l))(l+\varepsilon-x)})$  is increasing in  $\varepsilon$ , and so to maintain the equality of the

left-hand side and right-hand side of Equation (A2) we must have that  $A_k(l) > A_k(l + \varepsilon)$ .

(c). First we note that given an arrival rate  $\lambda$ :

$$\Pr\{\bar{L}_{k-1} \leq l\} \geq \Pr\{\bar{L}_k \leq l\}, \quad (\text{A3})$$

because the sum of  $k + 1$  non-negative random variables will in probability be greater than the sum of  $k$  random variables.

From Equation (A3):

$$\begin{aligned} & \int_0^l B_{k-1}(x) (1 - e^{-(\mu - A_k(l))(l-x)}) dx \\ & \geq \int_0^l B_k(x) (1 - e^{-(\mu - A_k(l))(l-x)}) dx. \end{aligned}$$

Therefore, for a given  $\tau$ :

$$\begin{aligned} & \int_0^l B_{k-1}(x) (1 - e^{-(\mu - A_{k-1}(l))(l-x)}) dx, \\ & = \int_0^l B_k(x) (1 - e^{-(\mu - A_k(l))(l-x)}) dx, \end{aligned}$$

implies that:

$$\begin{aligned} & \int_0^l B_{k-1}(x) (1 - e^{-(\mu - A_{k-1}(l))(l-x)}) dx \\ & \leq \int_0^l B_{k-1}(x) (1 - e^{-(\mu - A_k(l))(l-x)}) dx, \end{aligned}$$

thus,  $A_{k-1}(l) \geq A_k(l)$ . ■

Recall at this point the condition for  $C_k$  that:

$$\frac{\alpha(l_k)}{\alpha(\bar{l}_k)} \geq \frac{1 - \alpha(\bar{l}_{k+1})\rho}{1 - \alpha(\bar{l}_k)\rho}.$$

**Proposition 1.** *For a given  $k$ ,  $R_k \leq R_{k+1}$  if and only if  $C_k$  holds.*

**Proof.** Let  $\pi_i^k$  be the fraction of time that the system under scenario  $S_k$  spends in state  $i$ . Thus,  $\pi_0^k$  is the fraction of time the system is empty so that  $\mu(1 - \pi_0^k)$  is the throughput. Thus:

$$R_k^0 \leq R_{k+1}^0 \text{ iff } \mu(1 - \pi_0^k) \leq \mu(1 - \pi_0^{k+1}) \text{ iff } \frac{1}{\pi_0^k} \leq \frac{1}{\pi_0^{k+1}}. \quad (\text{A4})$$

Now define:

$$Q_i \equiv \prod_{j=0}^i \frac{\lambda_j}{\mu}, \quad i = 0, 1, 2, \dots, \infty \quad \text{and} \quad Q_{-1} \equiv 1,$$

where  $\lambda_j = \lambda\alpha(l_j)$ ,  $j = 0, \dots, k-1$ , and  $\bar{\lambda}_k = \lambda\alpha(\bar{l}_k)$  and noting that  $\lambda_i = \bar{\lambda}_k$  for  $i \geq k$ . We can now write the

steady-state probabilities:

$$\pi_i^k = \begin{cases} \pi_0^k \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu} = \pi_0^k Q_{i-1} & \text{for } i \leq k, \\ (\pi_0^k Q_{k-1}) \left(\frac{\bar{\lambda}_k}{\mu}\right)^{i-k} & \text{for } i \geq k+1, \end{cases}$$

$$\Rightarrow \sum_{i=k+1}^{\infty} \pi_i^k = (\pi_0^k Q_{k-1}) \sum_{i=k+1}^{\infty} \left(\frac{\bar{\lambda}_k}{\mu}\right)^{i-k},$$

$$\Rightarrow \pi_k^k + \sum_{i=k+1}^{\infty} \pi_i^k = (\pi_0^k Q_{k-1}) \left(\frac{\mu}{\mu - \bar{\lambda}_k}\right).$$

Recalling that  $1 = \sum_{i=0}^{\infty} \pi_i^k$  we have that:

$$\frac{1}{\pi_0^k} = \sum_{j=-1}^{k-2} Q_j + Q_{k-1} \left(\frac{\mu}{\mu - \bar{\lambda}_k}\right).$$

So Equation (A4) holds if and only if:

$$Q_{k-1} \left(\frac{\mu}{\mu - \bar{\lambda}_k}\right) \leq Q_{k-1} \left(1 + \frac{\lambda_k}{\mu} \left(\frac{\mu}{\mu - \bar{\lambda}_{k+1}}\right)\right),$$

if and only if:

$$\frac{\mu}{\mu - \bar{\lambda}_k} \leq \frac{\mu - \bar{\lambda}_{k+1} + \lambda_k}{\mu - \bar{\lambda}_{k+1}}.$$

Cross-multiplying and canceling terms yields:

$$\frac{\lambda_k}{\bar{\lambda}_k} \geq \frac{\mu - \bar{\lambda}_{k+1}}{\mu - \bar{\lambda}_k}$$

or

$$\frac{\alpha(\bar{l}_k)}{\alpha(\bar{l}_k)} \geq \frac{\mu - \alpha(\bar{l}_{k+1})\lambda}{\mu - \alpha(\bar{l}_k)\lambda}.$$

The result follows by dividing the right-hand side by  $\mu$  on top and bottom. ■

**Proposition 2.**  $W_0 \geq W_1$  and  $V_0 \geq V_1$ .

Before we prove the two statements of the proposition, we prove the following two lemmas:

**Lemma A1.**  $W_0 \geq W_1$  iff  $\bar{\lambda}_0 \geq \bar{\lambda}_1$  where  $W_0$  and  $W_1$  refer to the waiting times for scenario  $S_0$ , with submission rate  $\bar{\lambda}_0$ , and scenario  $S_1$  with arrival rates  $\lambda_0$  and  $\bar{\lambda}_1$ .

**Proof.** For scenario  $S_0$  the average waiting time is exactly the formula for the average wait in an M/M/1 queue with an arrival rate of  $\bar{\lambda}_0$ , thus  $W_0 = 1/(\mu - \bar{\lambda}_0)$ . For scenario  $S_1$  the average wait given you are in state 0 is  $1/\mu$ , and the average wait in state 1 is the time to leave state 1 to go to state 0, which is exactly the average wait of an M/M/1 queue with an arrival rate of  $\bar{\lambda}_1$ , namely

$1/(\mu - \bar{\lambda}_1)$ . If we now weight these times by the percentage of times we will incur them we have for system time:

$$W_1 = \frac{1}{\lambda_0\pi_0 + \bar{\lambda}_1\bar{\pi}_1} \left( \left(\frac{\lambda_0\pi_0}{\mu}\right) + \left(\frac{\bar{\lambda}_1\bar{\pi}_1}{\mu} + \frac{\bar{\lambda}_1\bar{\pi}_1}{\mu - \bar{\lambda}_1}\right) \right)$$

$$= \frac{1}{\mu} + \frac{\bar{\lambda}_1\bar{\pi}_1}{\lambda_0\pi_0 + \bar{\lambda}_1\bar{\pi}_1} \left(\frac{1}{\mu - \bar{\lambda}_1}\right)$$

$$= \frac{1}{\mu} \left(1 + \frac{\bar{\lambda}_1}{\mu - \bar{\lambda}_1}\right)$$

$$= \frac{1}{\mu - \bar{\lambda}_1}.$$

The next to last equality follows because the effective or time averaged arrival rate is:  $\lambda_0\pi_0 + \bar{\lambda}_1\bar{\pi}_1 = \mu(1 - \pi_0) = \mu(\bar{\pi}_1)$  so  $\bar{\pi}_1/(\lambda_0\pi_0 + \bar{\lambda}_1\bar{\pi}_1) = 1/\mu$ . Therefore,  $W_0 \geq W_1$  if and only if  $\bar{\lambda}_0 \geq \bar{\lambda}_1$ . ■

**Lemma A2.**  $V_0 \geq V_1$  if  $\mu \geq \bar{\lambda}_1$  and  $\lambda_0 \geq \bar{\lambda}_1$ , where  $V_0$  and  $V_1$  refer to the variances of waiting times for scenarios  $S_0$ , with arrival rate  $\bar{\lambda}_0$ , and  $S_1$  with arrival rates  $\lambda_0$  and  $\bar{\lambda}_1$ .

**Proof.** We first define  $X$  as the random variable for waiting time, and  $Y$  as the random variable for the state of the system:

$$V_0 = \left(\frac{1}{\mu - \bar{\lambda}_0}\right)^2.$$

To compute  $V_1$ , we use the identity:

$$\text{Var}[X] = \text{Var}_Y [E[X|Y]] + E_Y [\text{Var}[X|Y]].$$

For scenario  $S_1$ :

$$E[X|Y=0] = \frac{1}{\mu} \text{ and } E[X|Y=1] = \frac{1}{\mu} + \frac{1}{\mu - \bar{\lambda}_1},$$

while

$$\text{Var}[X|Y=0] = \frac{1}{\mu^2} \text{ and}$$

$$\text{Var}[X|Y=1] = \frac{1}{\mu^2} + \frac{1}{(\mu - \bar{\lambda}_1)^2}.$$

Therefore

$$V_1 = [\lambda_0\mu(2(\lambda_0^2 - \bar{\lambda}_1^2)^2 + 2(\lambda_0 - \bar{\lambda}_1)(\lambda_0 + \bar{\lambda}_1)(\lambda_0 + 2\bar{\lambda}_1)\mu + (\lambda_0^2 + 2\lambda_0\bar{\lambda}_1 + 2\bar{\lambda}_1^2)\mu^2)]$$

$$\times / [(\lambda_0 + \bar{\lambda}_1)^3(\bar{\lambda}_1 - \mu)^2(\lambda_0 - \bar{\lambda}_1 + \mu)^3].$$

Because

$$V_0 = \frac{1}{(\mu - \bar{\lambda}_0)^2} \geq \frac{1}{(\mu - \bar{\lambda}_1)^2} \text{ (because } \bar{\lambda}_1 \leq \bar{\lambda}_0),$$

it is sufficient to prove that  $V_1 \leq 1/(\mu - \bar{\lambda}_1)^2$ . This is

equivalent to proving that:

$$H(\lambda_0, \bar{\lambda}_1, \mu) = (\lambda_0 + \bar{\lambda}_1)^3(\lambda_0 + \bar{\lambda}_1 + \mu)^3 - \lambda_0\mu(2(\lambda_0 - \bar{\lambda}_1^2))^2 + 2(\lambda_0 - \bar{\lambda}_1)(\lambda_0 + \bar{\lambda}_1)(\lambda_0 + 2\bar{\lambda}_1)\mu + (\lambda_0^2 + 2\lambda_0\bar{\lambda}_1 + 2\bar{\lambda}_1^2)\mu^2 \geq 0,$$

$$\frac{\partial H}{\partial \lambda_0} = 6\lambda_0^5 + 5\lambda_0^4\mu + \bar{\lambda}_1^2(\bar{\lambda}_1 - \mu)^2\mu + 6\bar{\lambda}_0^2\bar{\lambda}_1\mu(\mu - \bar{\lambda}_1) + 4\lambda_0^3(\mu^2 + 3\bar{\lambda}_1\mu - 3\bar{\lambda}_1^2) + 2\lambda_0\bar{\lambda}_1(3\bar{\lambda}_1^3 + 2\bar{\lambda}_1\mu^2 + \mu^3 - 6\bar{\lambda}_1^2\mu).$$

It is easily shown that  $H(0, 0, \mu) = 0$ . Furthermore

$$\frac{\partial H}{\partial \lambda_0}(\lambda_0, \lambda_0, \mu) \geq 0 \quad \forall \lambda_0, \mu \geq 0.$$

Because  $\mu \geq \bar{\lambda}_1$  and  $\lambda_0 \geq \lambda_1$  the terms in the above expression for  $\partial H/\partial \lambda_0$  can be rearranged to show that:

$$\frac{\partial H}{\partial \lambda_0}(\lambda_0, \lambda_1, \mu) \geq 0 \quad \forall \lambda_0, \mu \geq 0.$$

The result follows. ■

**Proof of Proposition 2.** We can now complete the proof of Proposition 2 by showing that  $\bar{\lambda}_0 \geq \bar{\lambda}_1$ ,  $\mu \geq \bar{\lambda}_1$  by definition of  $\bar{\lambda}_1$  and  $\bar{\lambda}_0 \geq \bar{\lambda}_1$  if and only if  $\lambda\alpha(\bar{l}_0) \geq \lambda\alpha(\bar{l}_1)$  which is true because  $\bar{l}_0 \leq \bar{l}_1$ . ■

### Appendix B

The following are the parameter values used in generating the numerical results displayed in Figs. 5-7.

Step function demand curve:

$$D(x) = \lambda\alpha(x) = \begin{cases} 1 & \text{if } x < 6, \\ 0 & \text{if } x \geq 6. \end{cases}$$

Exponential

demand curve :

$$D(x) = \lambda\alpha(x) = e^{-x/10},$$

$\mu = 1.5, \lambda = 1, \tau = 0.95.$

### Biographies

Gregory Dobson is an Associate Professor of Operations Management at the Simon School of Business, University of Rochester and earned his Ph. D. in Operations Research from Stanford University. He serves on the Editorial Boards of *M&SOM*, *Interfaces*, and *International Journal of Services and Operations Management*. His work has been published in *Operations Research*, *Management Science*, *Marketing Science*, *Transportation Science*, the *European Journal of Operational Research*, *IIE Transactions*, *Production and Operations Management*, and *Mathematics of Operations Research*.

Edieal J. Pinker is an Associate Professor of Computers and Information Systems at the Simon School of Business, University of Rochester and earned his M.S. and Ph.D. in Operations Research from the Massachusetts Institute of Technology. He serves on the Editorial Board of *Management Science*, *M&SOM*, *POMS*, *Decisions Science* and *IJOR*. His work has been published in *Management Science*, *Manufacturing and Service Operations Management*, the *European Journal of Operational Research*, *IIE Transactions* and the *Communications of the Association of Computing Machinery*.

Contributed by the Supply Chains/Production-Inventory Systems Department