IIE Transactions (2001) **33**, 1019–1029

Analytic models for when and how to expedite in make-to-order systems

HASAN ARSLAN[1], HAYRIYE AYHAN[2] and TAVA LENNON OLSEN[3]*

[1]*Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109-2117, USA*
E-mail: harslan@mlt.edu
[2]*School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA*
E-mail: hayhan@isye.gatech.edu
[3]*John M. Olin School of Business, Washington University in St. Louis, St. Louis, MO 63130-4899, USA*
E-mail: olsen@olin.wustl.edu

Received July 1999 and accepted May 2001

Expediting is defined as using overtime or subcontracting to supplement regular production. This is usually done when the number of backorders has grown to be unacceptably large. In this paper, we consider analytic models for deciding when and how to expedite in a single-product make-to-order environment. We derive the structure of the optimal expediting policy in both continuous- and discrete-time cases. The continuous-time model corresponds best to subcontracting and the discrete-time model corresponds to either overtime or subcontracting. Models for performance analysis of the continuous-time case are also given.

1. Introduction

In a firm where product variety is part of its competitive advantage, make-to-order production may be preferable to keeping inventory as is discussed, for example, by Arreola-Risa and DeCroix, (1998). However, today's competitive marketplace demands short lead times. Unless a firm has a lot of extra capacity lead times will tend to be variable and may be quite long. Therefore, firms often need to use expediting techniques such as subcontracting or overtime to reduce lead times. Expediting will have a cost associated with it but may be worthwhile to prevent excessive delays. This paper provides analytic models for deciding how and when to expedite in make-to-order systems.

We use the term overtime to refer to production that takes place outside of regular time. It can be scheduled at discrete intervals only (e.g., at the end of the day or at the end of the week) and no new work arrives during this off-period. In many companies e.g., General Motors stamping facilities, (Jordan, 1997), overtime can be scheduled with very little advance notice. The manager looks at the current workload near the end of the week and decides whether or not to schedule overtime. As overtime can be very lucrative for employees there is usually sufficient workforce willing to cover overtime production even at short notice.

We use the definition of subcontracting given in Bradley (1997), namely: "The procurement from another firm of a product that the OEM simultaneously produces." Therefore the only practical difference between overtime and subcontracting is that subcontracting can occur at any time and overtime is only possible at discrete intervals. Bradley (1997) provides motivation from the electronics industry for why companies use subcontracting. The specific motivation for subcontracting considered in this paper is that the manager is willing to pay the extra associated costs to get the system back into control.

We look at analytic models of expediting both in the continuous- and discrete-time contexts. The continuous-time context is reserved for subcontracting while the discrete-time context covers both subcontracting and overtime. We assume that production is make-to-order and no inventory is kept. We provide structural results for how to manage expediting. In the continuous-time context we provide performance analysis of systems with expediting.

A number of authors have provided models for overtime production. Overtime production in queueing networks has been considered by Karmarkar *et al.* (1987) and Bitran and Tirupati (1991). Both papers approximate overtime production by appropriately scaling processing time. Rubin and Robson (1990) consider overtime at the end of a period that can only be used to finish the service that is currently in progress. Dellaert and Melo (1998) consider lot sizing in make-to-order production systems with overtime and due dates. Overtime has also been

*Corresponding author

looked at from the deterministic scheduling perspective (see, for example, Akkan (1996) and the references therein).

Duenyas *et al*. (1993), Duenyas *et al*. (1997) and Hopp *et al*. (1993) consider overtime in the context of inventory systems with production quotas. Each paper provides structural results for a number of different models. Hopp *et al*. (1993) assume that the company can sell everything that it can make and therefore the goal is to always produce a fixed production quota. Duenyas *et al*. (1993) apply this work to CONWIP systems with a fixed cost of overtime. Duenyas *et al*. (1997) investigate the decision of setting the production quota for discrete-time models with and without backlogging. Once the quota has been set, the plant will try to meet this in each period. They also show that if the inventory level is less than s, the plant needs to use the safety capacity to raise the net inventory to S units. Otherwise the plant should not use the safety capacity at all.

There has been some related work on systems where delivery lead times can be changed. One of the earliest papers is by Fukuda (1964); in this paper he considers product delivery with negotiable lead times, where later deliveries are at a discounted cost. In the paper the author derives optimal policies under the condition that products can be delivered with a normal lead time, or with a one period delay. Daniel (1963) derives structural results for an n period inventory model where emergency orders arrive immediately but regular orders are delayed one period. Moinzadeh and Schmidt (1991) consider a system where inventory can be replenished through either a normal or a more expensive emergency resupply channel. They use PDEs to derive steady-state distributions under a reasonable heuristic policy.

Less work appears to have been done in the subcontracting arena. Van Mieghem (1999) looks at the interaction between capacity investment and subcontracting. He considers different contracts for subcontracting and examines whether such contracts co-ordinate the supply chain. Bradley (1998a) has considered subcontracting in a make-to-stock model with backorders in discrete-time. In each period, the following sequence is repeated: (a) demand is realized; (b) in-house and subcontracting amounts are determined and received; (c) demand is fulfilled; and (d) inventory and back-ordering costs are incurred. A stationary two-parameter base-stock policy is shown to be optimal for the infinite horizon discounted case and for the infinite horizon average-cost case where shortfall is bounded. This work is extended by Bradley (1997) and Bradley and Glynn (1999) where the relationship between capacity and subcontracting is explicitly considered. In addition these works contain an analysis of a Brownian model operating under the two-parameter base-stock policy. In parallel work to ours, Bradley (1998b) develops an $M/M/1$ model of subcontracting where make-to-stock production is allowed. One key difference from our model (other than the fact that his model is make-to-stock) is that the subcontracting facility is a parallel facility so that production occurs at rate $\beta + \gamma$ when subcontracting used and at rate β when it is not.

This paper is organized as follows. Section 2 outlines the model used. Section 3 finds the structure of the optimal policies for continuous-time models. Steady-state performance analysis under the optimal policy is also provided. Section 4 finds the structure of the optimal policies for discrete-time models. Finally, Section 5 concludes the paper.

2. Model description

We consider a single production system producing a single type of product. Production times are assumed to be independent and identically distributed with mean $1/\mu$. Interarrival times for orders are independent and identically distributed, are independent of the processing times, and have mean $1/\lambda$. Following standard queueing notation, a $G/M/1$ system refers to a system with exponential processing times, an $M/G/1$ system has exponential interarrival times, and an $M/M/1$ system has both exponential processing times and exponential interarrival times. Let $\rho = \lambda/\mu$. We assume that the management does not accept any more orders if the number of orders backlogged reaches a predetermined value N (a large integer). We use the word "backlogged" or "backordered" to refer to the number of orders on hand that have not been expedited. We do this to avoid double-counting because backlogging cost will be explicitly considered in the expediting cost for expedited orders.

Production is make-to-order and no inventory is kept in advance. Notice that this model incorporates systems where orders may be specialized and therefore the general service time may reflect the fact that different orders have different requirements. Expediting may occur in two modes. The first is continuous where expediting may occur at any time. The second is discrete where expediting can only occur at discrete-time periods. The former corresponds best to subcontracting while the later covers both subcontracting and overtime. We assume that any number of orders may be expedited and we receive what we have ordered after a (possibly zero) lead time L that is independent of the system state. Note that the lead time model does not contain any notion of congestion.

The cost of expediting is made up of a fixed cost K and a per unit cost p. There is a backorder penalty cost of b per unit per unit time. Models are considered both with respect to a discounted-cost criteria, where we use α for the discount rate in continuous-time models and β for the discount factor in discrete-time models, and with respect to a long-run average cost criteria. In the long-run average cost criteria, if i units are expedited then the expected cost is $K + i(p + bE[L])$. We let $c = p + bE[L]$ be the total per unit cost associated with expediting in the average

cost criteria model. Similarly, we have $c_\alpha = p + bE[e^{-\alpha L}\mathbb{1}(L > 0)]$ as the total per unit cost associated with expediting in continuous-time models with discounted-cost criteria where, throughout the paper,

$$\mathbb{1}(A) = \begin{cases} 1 & \text{if the event } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

Also

$$c_\beta = p + bE\left[\frac{1-\beta^L}{1-\beta}\right]$$

is the total per unit cost associated with expediting in discrete-time models with discounted-cost criteria.

The optimal policy will be shown to be a threshold-type policy where expediting occurs once the number of orders reaches S. However, expediting should only be used to bring the number of orders down to some fixed level s which may be greater than zero if the per unit charge for expediting is significant. If $c = 0$ then s must be equal to zero in all of the models because expediting one extra unit adds no extra cost but will decrease backlogging costs.

3. Continuous-time models

In this section, we focus on continuous-time models where expediting may take place at any time. We show that for both the $G/M/1$ and the $M/G/1$ models, under both an infinite horizon discounted-cost criteria and a long-run average cost criteria, the optimal policy is an (s, S) threshold-type policy (i.e., if the number of orders waiting to be processed is greater than or equal to S then management should expedite enough to reduce the number of orders backordered to s). We also consider the $M/M/1$ system separately. Clearly, the $M/M/1$ model is a special case of the $M/G/1$ and the $G/M/1$ models. However the analysis is much simpler for the $M/M/1$ system because uniformization can be used to convert the continuous-time problem to an equivalent discrete-time problem. We provide this alternative discrete-time formulation for the $M/M/1$ system.

We model the problem of finding the optimal expediting policy as a semi-Markov decision process. In the $G/M/1$ system, we assume that the decision maker observes the system at the time of new order arrivals. Depending on the number of units backordered, he/she decides whether or not to use expediting. If expediting is used, he/she also determines how much to order. On the other hand, in the $M/G/1$ system, the decision maker observes the system at the time of service completions. He/she again decides whether or not to use expediting and, if expediting is used, the amount to order. In the $M/M/1$ system, both order arrival times and service completion times are decision epochs.

This section is organized as follows. Sections 3.1 and 3.2 provide structural results with respect to a discounted-cost criteria and a long-run average cost criteria, respectively. Section 3.3 provides steady-state performance analysis for both the $G/M/1$ and the $M/G/1$ models, as well as the special case of a $M/M/1$ system, operating under the optimal policy.

3.1. *Infinite horizon discounted-cost criteria*

This section considers the minimization of the infinite horizon discounted-cost for $G/M/1$, $M/G/1$ and $M/M/1$ systems.

3.1.1. *The $G/M/1$* system

Suppose that the processing times of orders are independent exponential random variables with rate μ and the times between order arrivals are independent identically distributed continuous random variables with common cumulative distribution function $F(\cdot)$ and mean $1/\lambda$. It is assumed that $F(0) < 1$. Since the expediting decision is based on the number of units backordered, we will capture the state of the system by the number of orders waiting (including the one which has just arrived) at the time of a new arrival. We will use the negative integers to denote our state space E. Thus, $E = \{-1, -2, \ldots, -N\}$ and at the time of an observation if the system state is i, this implies that $-i$ orders are backordered. The reason behind using the negative integers to denote the state space is simply to facilitate the characterization of the optimal policy. As is shown below, with this definition of the state space the equations that yield the optimal policies look very similar to those that appear in well-studied inventory models. Let $v(i)$ be the optimal value function of the infinite horizon discounted-cost problem given that the initial state is i. It is well-known that equations of the following form (*optimality equations*) characterize values and optimal policies in infinite horizon models (see, for example, Puterman, 1994)

$$v(i) = \min_{i \le a \le 0} \{K\mathbb{1}(a - i > 0) + c_\alpha(a - i) + g(a) + \sum_{j=-\infty}^{0} p_{aj} v(\max\{j-1, -N\})\} \quad \forall\, i \in E, \qquad (1)$$

where a is the action taken in state i representing $-1 \times$ (number of orders backordered after expediting), i.e., $-a$ is the number of orders left after expediting has occurred,

$$g(a) = E\left[b\int_0^{\tau_1} e^{-\alpha t} \max\{-a - M(t), 0\}\mathrm{d}t\right] = \int_0^\infty \int_0^x e^{-\alpha t} \sum_{k=0}^{-a} b(-a-k)\frac{e^{-\mu t}(\mu t)^k}{k!}\mathrm{d}t\,\mathrm{d}F(x),$$

where τ_1 is an interarrival time and $M(t)$ is the number of service completions in t time units and

$$p_{aj} = \begin{cases} 0 & \text{if } j < a, \\ \int_0^\infty e^{-\alpha t}\frac{e^{-\mu t}(\mu t)^{j-a}}{(j-a)!}\mathrm{d}F(t) & \text{if } a \le j < 0, \\ \int_o^\infty e^{-\alpha t}\mathrm{d}F(t) - \sum_{j=a}^{-1} p_{aj} & \text{if } j = 0. \end{cases}$$

Thus, p_{aj} is the probability of processing $j-a$ orders in between two consecutive order arrivals. It follows from Theorem 11.3.2 of Puterman (1994) that there exists a unique solution $v^* \in \mathbb{R}^\infty$ to the optimality equations in (1) and there exists a stationary deterministic optimal policy that yields this v^*. We will use $a^*(i)$ to denote the action that attains the minimum in (1) for state i.

The structural form of the optimal policy is described by the following theorem.

Theorem 1. *There exist non-negative integers s and S ($s < S$) such that if at the time of a new order arrival the number of orders backordered (including the one that has just arrived) is equal to* S, *then the decision maker should expedite enough to reduce the number of orders backordered to s. If the number of orders backordered is less than S, then the decision maker should not expedite.*

Proof. We can use a value iteration algorithm (see, for example, Puterman (1994)) to find the unique solution of the optimality equations (1). Starting with $v_0(i) = 0\ \forall i \in E$, we have the following recursive equations for the value iteration

$$v_n(i) = \min_{i \le a_n \le 0} \{K\mathbb{1}(a_n - i > 0) + c_\alpha(a_n - i) + g(a_n) + \sum_{j=-\infty}^{0} p_{a_n j} v_{n-1}(\max\{j-1, -N\})\} \quad \forall i \in E,$$

where a_n is the action taken at the nth decision epoch in state i. Since $g(a_n)$ is convex, it is straightforward to show inductively that $v_n(i)$ is K-convex. The proof of K-convexity by induction is analogous to Zabel's (1962) and therefore it is omitted. We also know that $\lim_{n\to\infty} v_n(i) = v^*(i)$, for all $i \in E$. Then it follows from Theorem 3 of Iglehart (1963) that $v^*(i)$ is also K-convex. Thus there exist integers $-s$ and $-S$ such that

$$a^*(i) = \begin{cases} -s & \text{if } i \le -S, \\ i & \text{otherwise.} \end{cases}$$

However, since the decisions are made at the time of new order arrivals, i can never be less than $-S$ under this policy which completes the proof. ■

3.1.2. *The $M/G/1$ system*

We now assume that the processing times of orders are independent identically distributed random variables with common cumulative distribution function $G(\cdot)$ with $G(0) < 1$ and mean $1/\mu$. The times between order arrivals are independent identically distributed exponential random variables with rate λ. As mentioned above, in the $M/G/1$ system the decision maker observes the state of the system at the time of service completions. We will again capture the state of the system by the number of orders waiting at the time of the decision epochs and use the non-positive integers to denote the state space $\mathscr{E}$. Thus, $\mathscr{E} = \{0, -1, \cdots, -N\}$. Then the optimality equations can be written as

$$v(i) = \min_{i \le a \le 0} \left\{K\mathbb{1}(a - i > 0) + c_\alpha(a - i) + g(a) + \sum_{j=-\infty}^{0} p_{aj} v(\max\{j, -N\})\right\} \quad \forall i \in \mathscr{E}, \qquad (2)$$

where

$$g(0) = E\left[e^{-\alpha\tau_1}\int_0^{S_1} e^{-\alpha t} b(\min\{1 + R(t), N\})\mathrm{d}t\right] = E[e^{-\alpha\tau_1}]\int_0^\infty \int_0^x e^{-\alpha t} b\left(\sum_{k=0}^{N-1}(1+k)\frac{e^{-\lambda t}(\lambda t)^k}{k!} + \sum_{k=N}^{\infty} N\frac{e^{-\lambda t}(\lambda t)^k}{k!}\right)\mathrm{d}t\mathrm{d}G(x),$$

and for $a < 0$

$$g(a) = E\left[\int_0^{S_1} e^{-\alpha t} b(\min\{-a + R(t), N\})\mathrm{d}t\right] = \int_0^\infty \int_0^x e^{-\alpha t} b\left(\sum_{k=0}^{N+a}(-a+k)\frac{e^{-\lambda t}(\lambda t)^k}{k!} + \sum_{k=N+a+1}^{\infty} N\frac{e^{-\lambda t}(\lambda t)^k}{k!}\right)\mathrm{d}t\mathrm{d}G(x),$$

where τ_1 is an interarrival time as before, S_1 is a service time, and $R(t)$ is the number of arrivals in t time units. Note that $g(\cdot)$ is quasi-convex. Furthermore

$$p_{0j} = \begin{cases} \int_0^\infty E[e^{-\alpha(\tau_1 + t)}]\frac{e^{-\lambda t}(\lambda t)^{-j}}{(-j)!}\mathrm{d}G(t) & \text{if } j \le 0, \\ 0 & \text{otherwise,} \end{cases}$$

and for $a < 0$

$$p_{aj} = \begin{cases} \int_0^\infty e^{-\alpha t}\frac{e^{-\lambda t}(\lambda t)^{a+1-j}}{(a+1-j)!}\mathrm{d}G(t) & \text{if } j \le a+1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, p_{aj} is the probability of having $a-j$ new order arrivals in an order processing time. Note that when $a = 0$, one has to wait until the next order arrival to start the process. Since (2) is similar to (1), the following result which says that the structure of the optimal expediting policy for the $M/G/1$ system is similar to that of the $G/M/1$ system is not surprising.

Theorem 2. *There exist non-negative integers s and S ($s < S$) such that if at the time of a service completion the number of units backordered is greater than or equal to S, then the decision maker should expedite enough to reduce the number of units backordered to s. If the number of units*

backordered is less than S, then the decision maker should not expedite.

Proof. Rewriting the optimality equations in (2) we have

$$v(i) = \min_{i \le a \le 0} \{K\mathbb{1}(a - i > 0) + c_\alpha a + g(a) + \sum_{j=-\infty}^{0} p_{aj} v(\max\{j, -N\})\} - c_\alpha i \quad \forall i \in \mathscr{E}. \quad (3)$$

Since $c_\alpha i$ does not affect the solution of the optimality equations, Equation (3) has the same structure as Equation (12) in Zheng (1991). Note that we can choose $c_\alpha a_n + g(a_n)$ as the G_α of Zheng (1991) and the summation $\sum_{l=0}^{\infty} p_l f_\alpha(j-l)$ of Zheng (1991) can be rewritten as $\sum_{-\infty}^{l=j} p_{j-l} f_\alpha(l)$ (i.e., dependence of the transition probabilities on the action is implicit in his expression). It then follows from Theorem 1 of Zheng (1991) that there exist integers $-s$ and $-S$ such that

$$a^*(i) = \begin{cases} -s & \text{if } i \le -S, \\ i & \text{otherwise.} \end{cases}$$ ■

3.1.3. *The M/M/1 system*

For this special case, we assume that the times between order arrivals and service times are exponential random variables with means $1/\lambda$ and $1/\mu$, respectively. Hence, we can apply uniformization techniques originally developed by Lippman (1975) and use discrete-time methods to characterize the structure of the optimal expediting policy. To this end, let $\Lambda = \lambda + \mu$ be the uniformization constant. The state of the system is again captured by the number of orders backlogged and non-positive integers are used to denote the state space $\mathscr{E} = \{0, -1, \ldots, -N\}$. The optimality equations for the uniformized model can be written as

$$v(i) = \min_{i \le a \le 0} \left\{ K\mathbb{1}(a - i > 0) + c_\alpha(a - i) + g(a) + \frac{\Lambda}{\Lambda + \alpha} \sum_{j=-N}^{0} p_{aj} v(j) \right\} \quad \forall i \in \mathscr{E},$$

where

$$g(a) = \frac{-ba}{\alpha + \Lambda}.$$

If $a = 0$

$$p_{0j} = \begin{cases} \frac{\lambda}{\Lambda} & \text{if } j = -1, \\ \frac{\mu}{\Lambda} & \text{if } j = 0, \\ 0 & \text{otherwise.} \end{cases}$$

If $-N + 1 \le a < 0$

$$p_{aj} = \begin{cases} \frac{\lambda}{\Lambda} & \text{if } j = a - 1, \\ \frac{\mu}{\Lambda} & \text{if } j = a + 1, \\ 0 & \text{otherwise.} \end{cases}$$

If $a = -N$

$$p_{aj} = \begin{cases} \frac{\lambda}{\Lambda} & \text{if } j = -N, \\ \frac{\mu}{\Lambda} & \text{if } j = -N + 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then it again follows from Theorem 1 of Zheng (1991) (or from an argument similar to the proof of Theorem 1 above) that the optimal expediting policy for the $M/M/1$ system is the same as that of the $G/M/1$ model.

Theorem 3. *There exist non-negative integers s and S ($s < S$) such that if (at the time of a new order arrival or at the time of a service completion) the number of orders backordered is equal to S, then the decision maker should expedite enough to reduce the number of orders backordered to s. If the number of orders backordered is less than S, then the decision maker should not expedite.*

Remark 1. *Note that Theorem* 1 *can also be proven using Theorem* 1 *of Zheng* (1991). *However, we leave the proof as it is in order to expose the reader to a different methodology in the special case that $g(\cdot)$ is convex.*

3.2. *The long-run average cost criteria*

We now consider minimizing the long-run average cost for $G/M/1$, $M/G/1$, and $M/M/1$ systems. For the threshold policy (s, S) let $v_\alpha^{(s,S)}(i)$ be its expected infinite horizon total discounted-cost, $g^{(s,S)}(i)$ be its long-run average cost and $v_T^{(s,S)}(i)$ be its expected cost up to time T starting in state i. Since the Markov chain is a unichain (i.e., the Markov chain corresponding to every stationary policy has a single recurrent class and possibly empty set of transient states)

$$\lim_{T \to \infty} \frac{1}{T} v_T^{(s,S)}(i),$$

exists for all $i \in E$ as an immediate consequence of Proposition 11.4.1 and Proposition 11.4.7 of Puterman (1994). By a standard Abelian result (Property A-8 of Heyman and Sobel (1984)) it follows that $\lim_{\alpha \downarrow 0} \alpha v_\alpha^{(s,S)}(i)$ exists and

$$\lim_{\alpha \downarrow 0} \alpha v_\alpha^{(s,S)}(i) = \lim_{T \to \infty} \frac{1}{T} v_T^{(s,S)}(i) = g^{(s,S)}(i).$$

From this discussion we can immediately conclude that the structures of the optimal policies that minimize the long average cost for the $G/M/1$, $M/G/1$, and $M/M/1$ systems are the same as those given in Theorem 1, Theorem 2, and Theorem 3 respectively.

3.3. *Queueing models*

This section gives steady-state performance analysis for queueing systems operating under the optimal policies from the previous sections. In particular, for given values of s and S, Sections 3.3.1, 3.3.2, and 3.3.3 give long-run average performance analysis for $M/M/1$, $G/M/1$, and $M/G/1$ systems, respectively.

3.3.1. *Analyzing the system in the M/M/1 case*

In this system, expediting occurs as soon as the number of orders backlogged hits S at which time the system state is immediately brought down to s. Thus the possible state space is $\{0, 1, \ldots, S-1\}$. Let π_i be the steady-state probability that the system is in state i, $i = 0, 1, \ldots, S-1$. The steady-state balance equations for $s > 0$ are:

$$\begin{cases} \pi_0\lambda = \pi_1\mu, \\ \pi_{S-i}\lambda + \pi_{S-i}\mu = \pi_{S-(i+1)}\lambda + \pi_{S-(i-1)}\mu \\ \quad \text{for } 1 < i < s \text{ or } s < i < S, \\ \pi_s\lambda + \pi_s\mu = \pi_{s-1}\lambda + \pi_{s+1}\mu + \pi_{S-1}\lambda, \\ \pi_{S-1}\lambda + \pi_{S-1}\mu = \pi_{S-2}\lambda. \end{cases}$$

If $s = 0$ then the balance equations are as follows:

$$\begin{cases} \pi_0\lambda = \pi_{S-1}\lambda + \pi_1\mu, \\ \pi_{S-i}\lambda + \pi_{S-i}\mu = \pi_{S-(i+1)}\lambda + \pi_{S-(i-1)}\mu \quad \text{for } 1 < i < S, \\ \pi_{S-1}\lambda + \pi_{S-1}\mu = \pi_{S-2}\lambda. \end{cases}$$

Rewriting the balance equations yields

$$\begin{cases} \pi_{S-i} = \sum_{k=0}^{i-1} \rho^{-k}\pi_{S-1} = \frac{1-\rho^{-i}}{1-\rho^{-1}}\pi_{S-1} \quad \text{for } \rho \neq 1, \\ \quad 1 < i \leq S - s, \\ \pi_{S-i} = \sum_{k=0}^{i-1} \rho^{-k}\pi_{S-1} = i\pi_{S-1} \quad \text{for } \rho = 1, \\ \quad 1 < i \leq S - s, \\ \pi_{s-i} = \sum_{k=i}^{S-s+i-1} \rho^{-k}\pi_{S-1} = \frac{\rho^{-i}(1-\rho^{-(S-s)})}{1-\rho^{-1}}\pi_{S-1} \quad \text{for } \rho \neq 1, \\ \quad 1 \leq i \leq s, \\ \pi_{s-i} = \sum_{k=i}^{S-s+i-1} \rho^{-k}\pi_{S-1} = (S-s)\pi_{S-1} \quad \text{for } \rho = 1, \\ \quad 1 \leq i \leq s. \end{cases}$$

Solving for π_{S-1}, we have

$$\pi_{S-1} = \begin{cases} \frac{\rho^{S-1}(1-\rho)^2}{1-\rho^{S-s}-\rho^S(1-\rho)(S-s)} & \text{for } \rho \neq 1, \\ \frac{2}{(S-s)(S+s+1)} & \text{for } \rho = 1. \end{cases}$$

Defining X as the steady-state number of customers backlogged

$$E[X] = \sum_{n=0}^{S-1} n\pi_n,$$

and hence for $\rho \neq 1$

$$E[X] = \left(\frac{\rho(s(s-1) - S(S-1))}{2(1-\rho)} + \frac{\rho(s-S)}{(1-\rho)^2} + \frac{\rho^2(\rho^{-S} - \rho^{-s})}{(1-\rho)^3}\right)\pi_{S-1},$$

and for $\rho = 1$

$$E[X] = \frac{1}{6}((S^3 - S) - (s^3 - s))\pi_{S-1}.$$

Therefore, for any given pair (s, S), the long-run average cost per unit time can be explicitly computed as $bE[X] + (K + c(S-s))\lambda\pi_{S-1}$, where the formulae for $E[X]$ and π_{S-1} are given above. Unfortunately, this function is not generally convex in s or S and therefore the minimum cost pair would need to be searched for by enumeration. However, as the function is explicit, this is not computationally difficult and for all cases we tested took less than a second. Figures 1–3 show the sensitivity of the parameters (s, S) to the various system parameters.

Figure 4 examines the sensitivity of S to backorder cost. The total per unit expediting cost c is set to zero so that $s = 0$ is optimal. We have chosen $\rho = 0.75$ and $K = 10$, 30, and 50, as shown. As could be expected, the expediting level is seen to be decreasing in backorder cost.

Figure 5 examines the sensitivity of s and S to expediting cost. We have chosen $\rho = 0.75$, $K = 30$, and $b = 1$. As could be expected, the expediting level is seen to be increasing in expediting cost.

Figure 6 examines the sensitivity of s and S to ρ. We have chosen $K = 30$, $c = 5$, and $b = 1$. It can be seen that s decreases as ρ increases but S first decreases and then increases. Note that as ρ increases so does the rate of increase of the queue length; therefore, in two systems with identical (s, S), the expediting rate will be greater in the system with the higher value of ρ.

3.3.2. *Analyzing the system in the G/M/1 case*

In this section, service times are assumed to be distributed exponentially with mean $1/\mu$ and interarrival times are

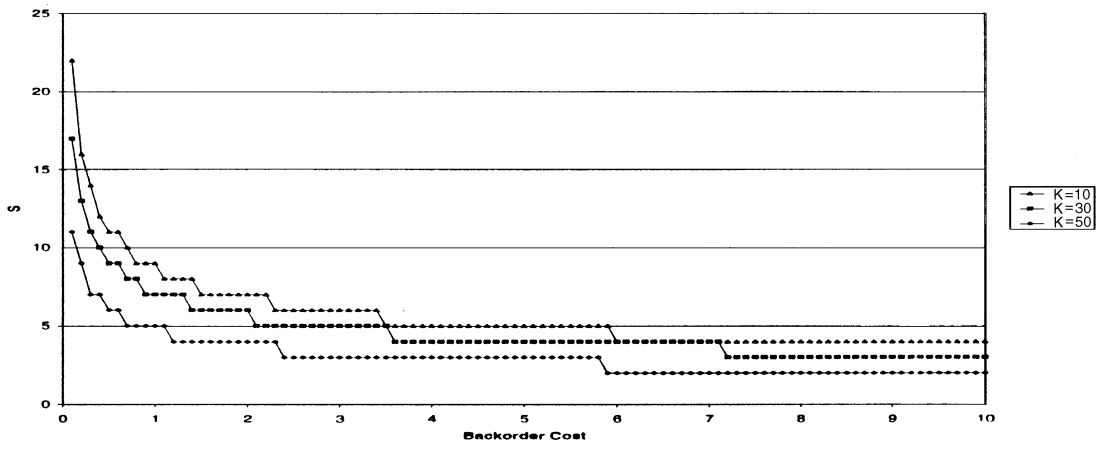

Fig. 1. Sensitivity of S to b when $c = 0$.

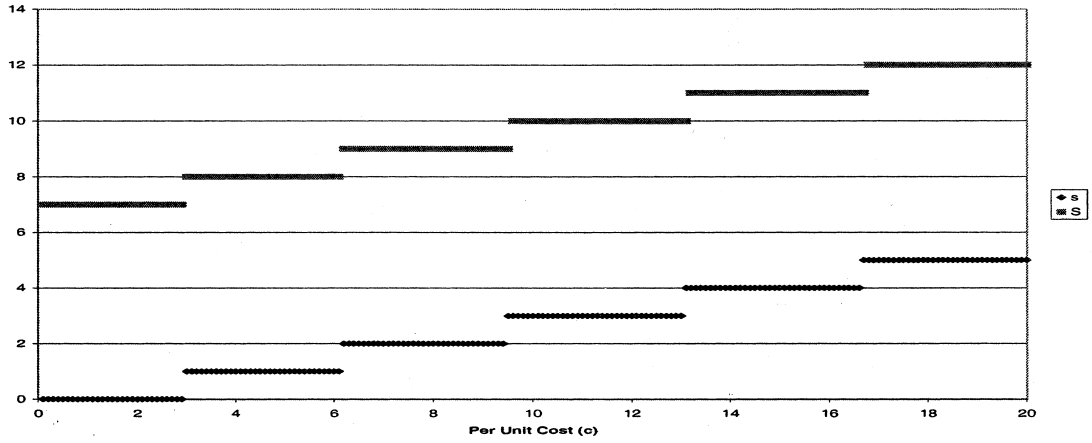

Fig. 2. Sensitivity of s and S to c.

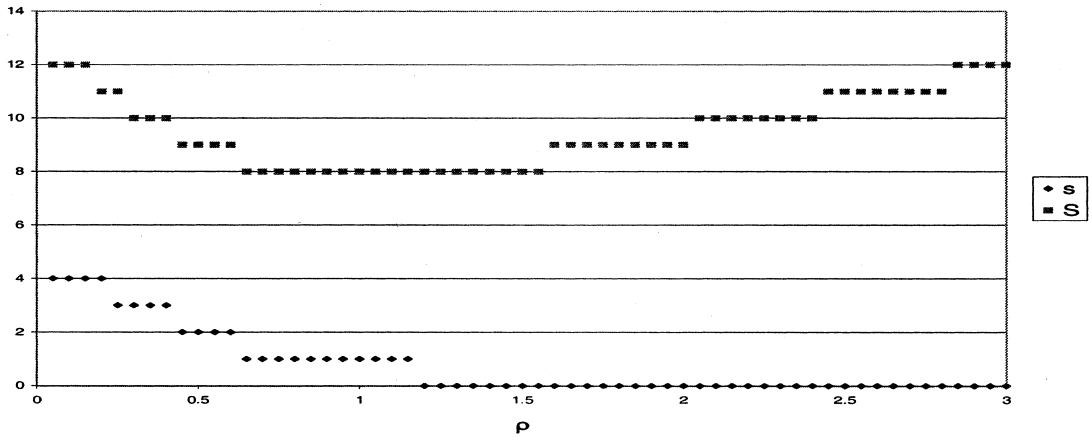

Fig. 3. Sensitivity of s and S to ρ.

assumed to have a general distribution F. For $n = 1, 2, 3, \ldots$, let X_n be the number of orders backlogged (including the one which has just arrived) at the time of the nth order arrival. Then $\{X_n\}$ is a discrete-time Markov chain on state space $1, \ldots, S$ with transition probability matrix $\boldsymbol{P}$ as follows:

$$\boldsymbol{P} = \begin{pmatrix} 1-p_0 & p_0 & 0 & 0 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & 0 \\ 1-p_0-p_1 & p_1 & p_0 & 0 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & 0 \\ 1-p_0-p_1-p_2 & p_2 & p_1 & p_0 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & 0 \\ 1-p_0-p_1-p_2-p_3 & p_3 & p_2 & p_1 & p_0 & \cdots & \cdots & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1-\sum_{i=0}^{S-2} p_i & p_{S-2} & p_{S-3} & p_{S-4} & p_{S-5} & \cdots & \cdots & p_{S-s-1} & p_{S-s-2} & \cdots & p_0 \\ 1-\sum_{i=0}^{s-1} p_i & p_{s-1} & p_{s-2} & p_{s-3} & p_{s-4} & \cdots & \cdots & p_0 & 0 & \cdots & 0 \end{pmatrix},$$

where p_n is the probability of having n departure events during an interarrival time, so that

$$p_n = \int_0^\infty \left(\frac{e^{-\mu t}(\mu t)^n}{n!} \right) \mathrm{d}F(t).$$

Rewriting $\boldsymbol{\pi} = \boldsymbol{\pi P}$ in component notation we have

$$\pi_1 = (1-p_0)\pi_1 + (1-p_0-p_1)\pi_2 + \ldots + \left(1 - \sum_{i=0}^{S-2} p_i\right)\pi_{S-1} + \left(1 - \sum_{i=0}^{s-1} p_i\right)\pi_S,$$

$$\pi_i = p_0\pi_{i-1} + p_1\pi_i + p_2\pi_{i+1} + \cdots + p_{S-i}\pi_{S-1} + p_{s-i+1}\pi_S, \quad 2 \le i \le s+1,$$

$$\pi_i = p_0\pi_{i-1} + p_1\pi_i + p_2\pi_{i+1} + \cdots + p_{S-i}\pi_{S-1}, \quad s+1 < i \le S-1,$$

$$\pi_S = p_0\pi_{S-1},$$

which may be solved explicitly by writing π_S in terms of π_{S-1}, π_{S-1} in terms of π_{S-2} (and lower order terms), and

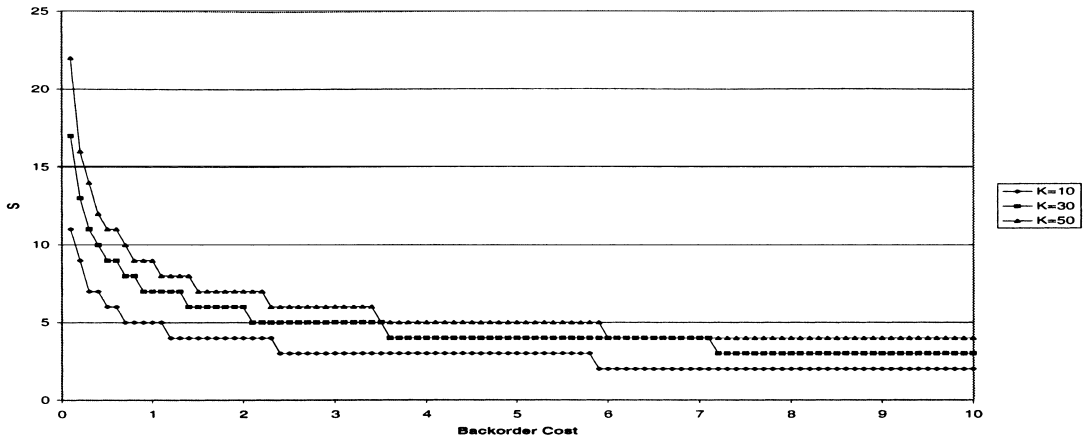

Fig. 4. Sensitivity of S to b when $c = 0$.

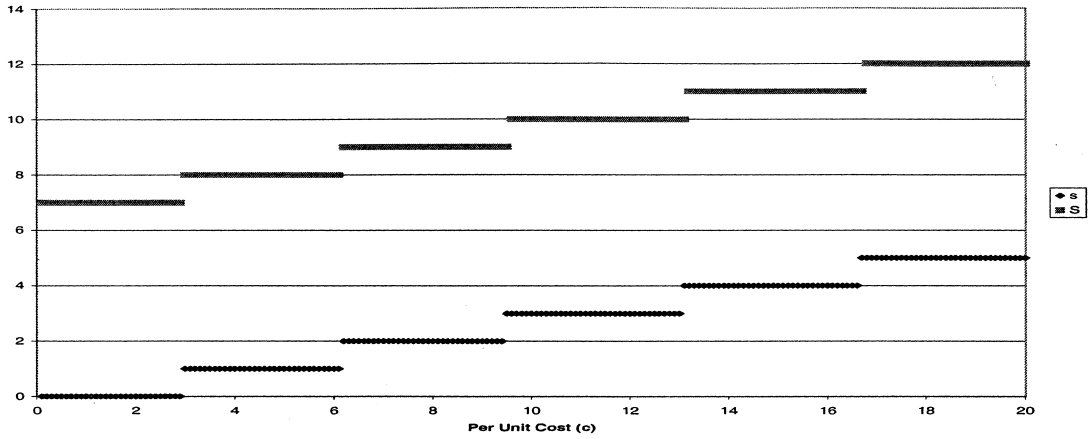

Fig. 5. Sensitivity of s and S to c.

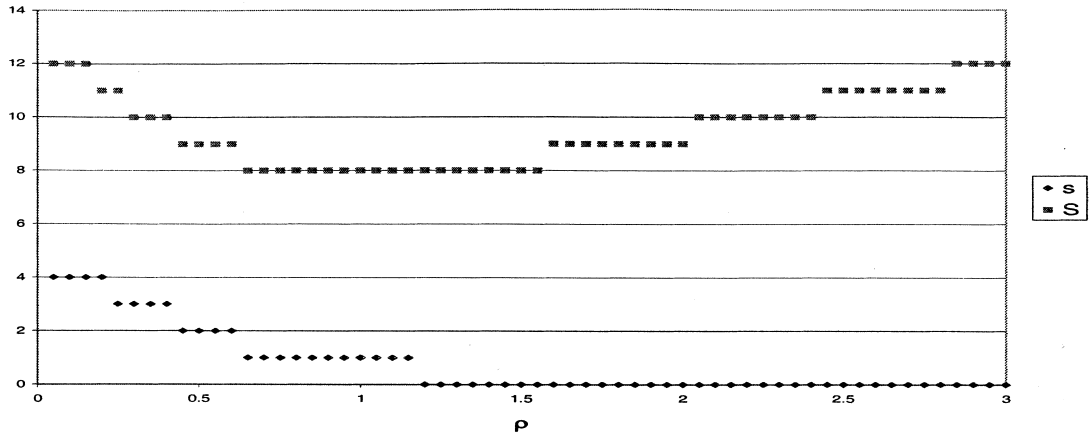

Fig. 6. Sensitivity of s and S to ρ.

so on. In practice however, as this is so notationally cumbersome, it is easier to solve these equations using a mathematical package such as Mathematica™.

If X is the steady-state number of customers backlogged then, using the fact that the expected time between transitions is $1/\lambda$, we have

$$E[X] = \sum_{i=1}^{S-1} \lambda m(i)\pi_i + \lambda m(s)\pi_S,$$

where

$$m(i) = \int_0^\infty \int_0^x \sum_{k=0}^{i} (i-k)\frac{e^{-\mu t}(\lambda t)^k}{k!}\mathrm{d}t\mathrm{d}F(x). \qquad (4)$$

is the expected number of customers in the system between transitions if the initial state is i. Therefore, for any given pair (s, S), the long-run average cost per unit time can be computed as $bE[X] + (K + c(S-s))\lambda\pi_S$.

3.3.3. *Analyzing the system in the M/G/1 case*

In this section, interarrival times are assumed to be exponentially distributed and service times are assumed to have a general distribution G. For $n = 1, 2, 3, \ldots$, let X_n be the number of customers left in the system following the nth service completion and possible expediting. We assume $S > 1$ so that expediting never occurs upon an arrival to an empty system. In the case $S = 1$, expediting occurs upon every arrival to the system, and therefore the long-run average cost per unit time can be computed as $\lambda(K + c)$. With $S > 1$, expediting occurs if there are S or more customers in the system upon a service completion therefore $\{X_n\}$ is a discrete-time Markov chain on state space $0, 1, \ldots, S-1$. It has transition probability matrix $\boldsymbol{P}$ as follows:

$$\begin{pmatrix}
q_0 & q_1 & q_2 & q_3 & q_4 & \ldots & \ldots & 1-\sum_{i=0}^{S-1} q_i + q_s & \ldots & \ldots & q_{S-1} \\
q_0 & q_1 & q_2 & q_3 & q_4 & \ldots & \ldots & 1-\sum_{i=0}^{S-1} q_i + q_s & \ldots & \ldots & q_{S-1} \\
0 & q_0 & q_1 & q_2 & q_3 & \ldots & \ldots & 1-\sum_{i=0}^{S-2} q_i + q_{s-1} & \ldots & \ldots & q_{S-2} \\
0 & 0 & q_0 & q_1 & q_2 & \ldots & \ldots & 1-\sum_{i=0}^{S-3} q_i + q_{s-2} & \ldots & \ldots & q_{S-3} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-q_0-q_1 & 0 & q_0 & q_1
\end{pmatrix},$$

where q_n is the probability of having n arrivals during a service time, so that

$$q_n = \int_0^\infty \left(\frac{e^{-\lambda t}(\lambda t)^n}{n!}\right)\mathrm{d}G(t).$$

Solving $\boldsymbol{\pi} = \boldsymbol{\pi P}$ for the stationary probability vector $\boldsymbol{\pi}$ yields:

$$\pi_i = q_i\pi_0 + \sum_{j=0}^{i} q_j\pi_{i-j+1}, \quad 0 \le i \le S-2, \quad i \ne s,$$

$$\pi_s = \left(1 - \sum_{i=0}^{S-1} q_i + q_s\right)\pi_0 + \sum_{j=1}^{s+1}\left(1 - \sum_{i=0}^{S-j} q_i + q_{s-j+1}\right)\pi_j + \sum_{j=s+2}^{S-1}\left(1 - \sum_{i=0}^{S-j} q_i\right)\pi_j,$$

$$\pi_{S-1} = q_{S-1}\pi_0 + \sum_{j=1}^{S-1} q_j\pi_{S-j}.$$

The above equations must be solved implicitly.

If X is the steady-state number of customers backlogged then

$$E[X] = \mu n(1)\pi_0 + \sum_{i=1}^{S-1} \mu n(i)\pi_i,$$

where

$$n(i) = \int_0^\infty \int_0^x \sum_{k=0}^{\infty} (i+k)\frac{e^{-\lambda t}(\lambda t)^k}{k!}\mathrm{d}t\mathrm{d}G(x). \qquad (5)$$

Note that the time between the n and $(n+1)$th departure from the system is distributed with mean $1/\mu$ unless $X_n = 0$, in which case it has mean $1/\lambda + 1/\mu$. We were able to use $1/\mu$ as the expected time between transitions in the above because the time the system is empty contributes zero to $E[X]$. The proportion of time spent in state $S-1$ is $(\pi_{S-1}/\mu)/(1/\mu + \pi_0/\lambda)$. Therefore, for any given pair (s, S), $S > 1$, the long-run average cost per unit time can be computed as $bE[X] + (K + c(S-s))\lambda^2\pi_{S-1}/(\lambda + \mu\pi_0)$.

4. Discrete-time models

In this section, we consider the discrete-time analog of the models considered in Section 3. We assume that the decision maker observes the system periodically at predetermined time epochs. Depending on the number of units backordered, he again determines whether or not to use expediting. Let D_n be the number of units demanded between the nth and the $(n+1)$th observation time. Similarly, let Y_n be the number of units produced between the nth and the $(n+1)$th observation time. Suppose

$$P\{D_n = k\} = q(k) \quad k = 0, 1, \ldots \forall n \ge 1,$$
$$P\{Y_n = k\} = f(k) \quad k = 0, 1, \ldots \forall n \ge 1.$$

We will again consider both the infinite horizon discounted-cost model and the long-run average cost model. The structure of the optimal policies is the same as the ones that appeared in Section 3.

4.1. *Infinite horizon discounted-cost criteria*

Since the expediting decision is based on the number of units backordered, we will again capture the state of the system by the number of orders waiting at the time of the observations. Thus, the state space is again $\mathscr{E} = \{0, -1, \ldots, -N\}$. A value iteration can again be used to find the structure of the optimal policy. With $v_0(i) = 0$ for all $i \in \mathscr{E}$, the recursive equations for the value iteration is given as

$$\begin{aligned} v_n(i) = \min_{i \le a \le 0} \{ & K\mathbb{1}(a - i > 0) + c_\beta(a - i) - ba \\ & + \beta \sum_{j=0}^{\infty} \sum_{k=j-a}^{\infty} f(k) q(j) v_{n-1}(0) + \\ & \beta \sum_{j=0}^{\infty} \sum_{k=0}^{j-a-1} f(k) q(j) v_{n-1}(\max\{a - j + k, -N\}) \} \\ & \forall i \in \mathscr{E}, \end{aligned} \tag{6}$$

where $0 \le \beta < 1$ is the discount factor.

Theorem 4. *There exist non-negative integers r and R $(r < R)$ such that if at the time of an observation the number of units backordered is greater than or equal to R, then the decision maker should expedite enough to reduce the number of units backordered to r. If the number of units backordered is less than R, then the decision maker should not expedite.*

Proof. Since the existence of a stationary optimal policy is guaranteed by Theorem 6.2.7 of Puterman (1994), the proof is analogous to the proof of Theorem 1 and it is omitted. ■

4.2. *Long-run average cost criteria*

We again assume that the state of the system is observed periodically and the state space is again $\mathscr{E}$. For the threshold policy (r, R) let $v_\beta^{(r,R)}(i)$ be its expected infinite horizon total discounted-cost, $g^{(r,R)}(i)$ be its expected long-run average cost and $v_n^{(r,R)}(i)$ be its expected cost up to time n starting in state i. Since the Markov chain has a finite state space it immediately follows from Proposition 8.1.1 of Puterman (1994) that

$$\lim_{n \to \infty} \frac{1}{n} v_n^{(r,R)}(i),$$

exists for all $i \in \mathscr{E}$. Then from Corollary 8.2.5 of Puterman (1994)

$$\lim_{\beta \uparrow 1} (1 - \beta) v_\beta^{(r,R)}(i) = \lim_{n \to \infty} \frac{1}{n} v_n^{(r,R)}(i) = g^{(r,R)}(i).$$

The following result is a direct consequence of the above discussion.

Theorem 5. *There exist non-negative integers r and R $(r < R)$ such that if at the time of an observation the number of units backordered is greater than or equal to R, then the decision maker should expedite enough to reduce the number of units backordered to r. If the number of units backordered is less than R, then the decision maker should not expedite.*

5. Conclusion

In this paper, we considered analytic models of expediting for single-server make-to-order production systems in both the continuous-time and discrete-time case. The optimal policy in all cases was shown to be a simple (s, S) type policy where expediting should occur once the number of units backlogged reaches (or exceeds) S. When expediting occurs it should bring the backlog down to s which may be non-zero if there is a per unit cost associated with expediting. Queueing models for evaluating specific pairs (s, S) were given for the continuous-time case under the long-run average cost criteria. Future work should look at modeling more complex queueing systems with expediting.

References

Akkan, C. (1996) Overtime scheduling: an application in finite-capacity real-time scheduling. *Journal of the Operational Research Society*, **47**, 1137–1149.

Arreola-Risa, A. and DeCroix, G.A. (1998) Make-to-order versus make-to-stock in a production-inventory system with general production times. *IIE Transactions*, **30**, 705–713.

Bitran, G.R. and Tirupati, D. (1991) Approximations for networks of queues with overtime. *Management Science*, **37**, 282–300.

Bradley, J.R. (1997) Managing assets and subcontracting policies. Ph.D. dissertation, Stanford University, Stanford, CA.

Bradley, J.R. (1998a) A discrete-time subcontracting model. Working Paper, Cornell University, Ithaca, NY.

Bradley, J.R. (1998b) Optimal control of an $M/M/1$ subcontracting model. Working Paper, Cornell University, Ithaca, NY.

Bradley, J.R. and Glynn, P.W. (1999) Managing the manufacturer-subcontractor relationship and the manufacturer's optimal capacity, inventory, and subcontracting policies. Working Paper, Cornell University, Ithaca, NY.

Daniel, K.H. (1963) A delivery-lag inventory model with emergency, in *Multistage Inventory Models and Techniques*, Scarf, H.E., Gilford, D.M. and Shelly, M.W. (eds,).

Dellaert, N.P. and Melo, M.T. (1998) Make-to-order policies for a stochastic lot-sizing problem using overtime. *International Journal of Production Economics*, **56–57**, 79–97.

Duenyas, I., Hopp, W.J. and Bassok, Y. (1997) Production quotas as bounds on interplant JIT contracts. *Management Science*, **43**, 1372–1386.

Duenyas, I., Hopp, W.J. and Spearman, M.L. (1993) Characterizing the output process of a CONWIP line with deterministic

processing and random outages. *Management Science*, **39**, 975–988.
Fukuda, Y. (1964) Optimal policies for the inventory problem with negotiable leadtime. *Management Science*, **10**, 690–708.
Heyman, D.P. and Sobel, M.J. (1984) *Stochastic Models in Operations Research*, Vol. I, McGraw Hill, New York.
Hopp, W.J., Spearman, M.L. and Duenyas, I. (1993) Economic production quotas for pull manufacturing systems. *IIE Transactions*, **25**, 71–79.
Iglehart, D.L. (1963) Optimality of (*s*,*S*) policies in the infinite horizon dynamic inventory problem. *Management Science*, **9**, 259–267.
Jordan, W. (1997) Personal communication.
Karmarkar, U.S., Kekre, S. and Kekre, S. (1987) Capacity analysis of a manufacturing cell. *Journal of Manufacturing Systems*, **6**, 165–175.
Lippman, S.A. (1975) Applying a new device in the optimization of exponential queueing system. *Operations Research*, **23**, 687–710.
Moinzadeh, K. and Schmidt, C.P. (1991) An (*S*-1,*S*) inventory system with emergency orders. *Operations Research*, **39**, 308–321.
Puterman, M.L. (1994) *Markov Decision Processes*, John Wiley, New York.
Rubin, G. and Robson, D.S. (1990) A single server queue with random arrivals and balking: confidence interval estimation. *Queueing Systems*, **7**, 283–306.
Van Mieghem, J.A. (1999) Coordinating investment, production and subcontracting. *Management Science*, **45**, 954–971.
Zabel, E. (1962) A note on the optimality of (*s*,*S*) policies in inventory theory. *Management Science*, **9**, 123–125.
Zheng, Y-S. (1991) A simple proof for optimality of (*s*,*S*) policies in infinite-horizon inventory systems. *Journal of Applied Probability*, **28**, 802–810.

Biographies

Hasan Arslan is a Ph.D. student in the Sloan School of Management at Massachusetts Institute of Technology. He completed the work in this paper while an undergraduate student in the Department of Industrial and Operations Engineering at the University of Michigan.

Hayriye Ayhan is an Assistant Professor in the School of Industrial and Systems Engineering at Georgia Institute of Technology. Dr. Ayhan received her B.S. in Industrial Engineering from Bogazici University, Turkey. She earned her M.S. and Ph.D. in Industrial Engineering from Texas A&M University. Her research interests lie in the area of analysis and control of queueing systems (in particular those that arise in manufacturing settings). Dr. Ayhan's research papers have been accepted for publication in *Operations Research, IEEE Transactions on Automatic Control, Queueing Systems, Journal of Applied Probability* and several other journals. She is a member of Institute for Operations Research and Management Science (INFORMS) and she is the secretary and treasurer of the INFORMS Applied Probability Society.

Tava Lennon Olsen is an Associate Professor in the Olin School of Business at Washington University in St. Louis. She was previously an Assistant Professor in the Department of Industrial and Operations Engineering at the University of Michigan. Dr. Olsen received her B.Sc. (honours) in Mathematics in 1990 from the University of Auckland, New Zealand. She earned both her M.S. in Statistics in 1992 and her Ph.D. in Operations Research in 1994 from Stanford University. Dr. Olsen's research interests include the stochastic modeling of manufacturing systems, supply-chain management, queueing systems, and applied probability. Among other journals, her publications have appeared in *Management Science, Operations Research,* and *IEEE Transactions on Automatic Control*. Dr. Olsen teaches courses in operations management. At Michigan she taught courses in simulation and stochastic processes and twice won the departmental teaching award. Dr. Olsen is currently an Associate Editor for *Management Science*.

Contributed by the Production Planning Department