Static and Dynamic Pricing Of Excess Capacity in a Make-To-Order Environment

Joseph M. Hall

Praveen K. Kopalle

David F. Pyke

Tuck School of Business at Dartmouth
100 Tuck Hall, HB 9000
Hanover, NH 03755

Phone: 603-646-0778
Fax: 603-646-1308

Joseph.M.Hall@Dartmouth.EDU
Praveen.K.Kopalle@Dartmouth.EDU
David.F.Pyke@Dartmouth.EDU

Static and Dynamic Pricing Of Excess Capacity in a Make-To-Order Environment

ABSTRACT

Recent years have seen advances in research and management practice in the area of pricing, and particularly in dynamic pricing and revenue management. At the same time, researchers and managers have made dramatic improvements in production and supply chain management. The interactions between pricing and production/supply chain performance, however, are not as well understood. Can a firm benefit from knowing the status of the supply chain or production facility when making pricing decisions? How much can be gained if pricing decisions explicitly and optimally account for this status? This paper addresses these questions by examining a make-to-order manufacturer that serves two customer classes – core customers who pay a fixed negotiated price and are guaranteed job acceptance, and "fill-in" customers who make job submittal decisions based on the instantaneous price set by the firm for such orders. We examine four pricing policies that span a range of complexity and required knowledge about the status of the production system at the manufacturer, including the optimal policy of setting a different price for each possible state of the queue. We demonstrate properties of the optimal policy, and we illustrate numerically the financial gains a firm can achieve by following this policy vs. simpler pricing policies. The four policies we consider are (1) state-independent (static) pricing, (2) allowing fill-in orders only when the system is idle, (3) setting a uniform price up to a cut-off state, and (4) general state-dependent pricing. Although general state-dependent pricing is optimal in this setting, we find that charging a uniform price up to a cut-off state performs quite well in many settings and presents an attractive trade-off between ease of implementation and profitability. Thus, a fairly simple heuristic policy may actually out-perform the optimal policy when costs of design and implementation are taken into account.

1. INTRODUCTION

MetalFab, Inc.[1] produces fabricated metal parts mostly for use in the power generation industry. The parts are made from expensive materials – some 4x8-foot sheets of material cost $20,000 – and, not surprisingly, the fabricated parts have very tight tolerances. MetalFab is a large job shop with about 60 highly skilled shop floor employees who operate metal bending and metal cutting machines, as well as a variety of welding equipment. At this writing, approximately 80% of MetalFab's output is sold directly to General Electric, or to first tier GE suppliers. In keeping with MetalFab's policy, we will refer to this output as belonging to GE.

MetalFab can forecast orders from GE and GE's suppliers, but the forecast error can be quite high. Sometimes MetalFab production planners will have firm forecasts – and these forecasts remain firm until the order is delivered. More often, however, GE will change the order quantity and due date several times while the order is outstanding. In fact, GE's systems will occasionally produce a purchase order that is already past due when the order is placed. (The authors observed a case in which an order placed in September had a due date of the previous May!) Taken together, the blend of firm forecasts, changes, and emergency orders create a situation that is well captured by a mean forecast with a fairly high variance around that mean. Similar situations could arise in cases when a single large customer aggregates demand forecasts from many different locations and provides an aggregate request to its supplier.

From MetalFab's perspective, GE orders form the core of its business, while the other orders that may take up the remaining 20% of its capacity are treated as "fill-in" orders. From a marketing standpoint, one approach to accepting and pricing fill-in orders is to take as many as possible, charge the same price as the core orders, and let the production planners and factory workers try to keep up. The danger with this approach, of course, is that it may not be a long-run profit-maximizing strategy; and service performance, for both GE and fill-ins, could suffer. An alternative approach is to proactively seek fill-in orders when capacity utilization is running low, charging a low price to attract those customers; and when the capacity utilization is high, charging a high price and accepting only limited fill-in orders. This alternative raises the issue of how to (i) price dynamically over time depending on the state of the production system, (ii) endogenously determine "low" versus "high" capacity utilization, and (iii) incorporate core

customer arrivals (i.e., orders from core customers) in determining the pricing policy for fill-in customers. For example, when a potential fill-in customer asks for a bid for a given part, what price should MetalFab quote? Should that price depend on the current level of congestion at the factory? If so, how? Finally, what benefits are available if the firm wisely uses capacity information when making pricing decisions? This paper addresses these questions.

We focus attention on four models that take into consideration both core and fill-in customer arrival rates: (1) state-independent (static) pricing – where MetalFab sets a price, p, for fill-in customers without regard to the current state of the factory; (2) allowing fill-in jobs at a chosen price p only when the factory is idle; (3) allowing fill-in jobs at a chosen price p only when there are s or fewer jobs in the production system, where both s and p are decision variables; and (4) general state dependent pricing – i.e. potentially setting a different price for fill-in orders for every possible state of the factory. To ensure satisfactory service, we impose a constraint on expected waiting time for core customers. We compare the optimal solutions obtained in the above four cases, report the magnitude of the benefit from utilizing increasing amounts of information, illustrate interesting properties of the solutions, and examine conditions under which one solution is superior to another.

Before reviewing the relevant literature, it is important to note that this problem is quite general and is generating much interest beyond high precision job shops like MetalFab. With the advent of modern pricing software such as that offered by DemandTec, ProfitLogic, and KhiMetrics, many companies now are devoting considerable time and energy to getting prices right. However, recent trade press articles suggest that firms have traditionally been slow to adopt sophisticated pricing models (Reda 2002), have priced products solely on cost (*At What Price? Guidelines for a Customer-Focused Pricing Strategy* 2000), and often simply employ "what-if" analyses without incorporating the interactions across functional areas (*Retail Revenue Management* 2001, Lester 2002). Further, the transition to the Euro has elevated this issue for companies doing business in Europe, and many are appointing senior "pricing officers" with direct responsibility over pricing decisions. Furthermore, many firms are beginning to realize that price changes should be made with a deeper understanding of the supply chain (Cisco Thought Leadership Summit 2001). If a firm cuts price to stimulate demand, but the factory or supply chain is currently overloaded, they risk some very unhappy customers. On the other

[1] The name of this privately owned company has been disguised at the owner's request.

hand, if the supply chain and factories currently have excess capacity, marketing personnel may wish to decrease price to consume some of that capacity. In addition, some of the leading suppliers of supply chain software are developing linkages to pricing software. Manugistics, for instance, has bought Talus, a revenue management software provider with the expressed intent of linking these two areas. This research is designed to generate insight for managers about the benefits of accounting for the supply chain when making pricing decisions.

The rest of this paper is organized as follows. In Section 2 we review the relevant literature. In Section 3, we present the four models and corresponding analytical results. A numerical comparison of the policies is presented in Section 4. Section 5 contains a summary discussion and directions for future research.

2. LITERATURE REVIEW

The last two decades have seen significant research progress on the interaction of pricing and operations. This literature falls into two fundamental categories: pricing/inventory models and pricing/queuing models. In the first case, generally speaking, prices are determined jointly with inventory decisions, or are determined based on current inventory levels. In the second case, prices are used to control the arrival rate to a queue or queues and may or may not be set based on the current queue length. For a recent review of this literature, see Fleischmann, Hall & Pyke (2004).

Pricing/inventory models have a long history, beginning with single period, single price, single quantity models like those of Whitin (1955), Karlin & Carr (1962), and Lau & Lau (1988). Multiperiod models typically assume a single, constant price, and deterministic demand. Examples include Wagner & Whitin (1958a), Wagner & Whitin (1958b), Thomas (1970), Kunreuther & Richard (1971), Kunreuther & Schrage (1973), Pekelman (1974), Eliashberg & Steinberg (1987), Eliashberg & Steinberg (1991), Gilbert (2000), Arvind Rajan & Steinberg (1992), and Sogomonian & Tang (1993).

Models with stochastic demand include Thomas (1974) who addresses an N period problem and proposes a heuristic policy of the form (s, S, p), where price is a parameter in the probability distribution of demand. Federgruen & Heching (1999) consider a infinite horizon, order-up-to model that has a stationary base stock policy as the optimal policy structure in the

case of zero leadtimes. When leadtimes are nonzero, the optimal policy depends on individual order epochs and becomes intractable. However, price is constant and there is no opportunity to change price within period. Chen & Simchi-Levi (2002) consider stationary infinite horizon models with fixed costs and show that the price p in the optimal policy depends on the inventory level. In a seasonal demand model, Biller, Chan, Simchi-Levi, & Swann (2002) and Chan, Simchi-Levi, & Swann (2001) show that, from an inventory perspective, pricing may help meet capacity restrictions by smoothing out demand. Our work differs from this stream of research insofar as we consider a make-to-order system and explicitly model the production system.

In the price/queuing literature, Naor (1969) suggests imposing tolls on new customers to keep queue size in check, where customers have Poisson arrivals and service times are exponential. Upon arrival, customers see the queue length and choose to stay or go. Balachandran (1972) studies a system in which customers make a payment to purchase a place in the queue. Larger payments buy higher priorities, and customers know the state of the system when they arrive. Customers have delay costs, so they minimize their total cost which includes payment and delay. Adiri & Yechiali (1974), on the other hand, examine the profit-maximizing price for jobs joining a priority queue where prices vary depending on the priority level of the queue. When a customer arrives, he receives information on the state of the system and then decides whether to join the queue or not. The authors use an iterative process where prices are set first, then customers calculate optimal control limits (i.e., join the queue if the queue length is $\leq n^*$), then the firm adjust prices, and so on. Stidham (1985) considers a variation where arrivals are restricted by charging a fee or closing the queue. He considers two cases; in one, the firm is ignorant of the current state of the system, in the other, the firm knows the current state and exercises dynamic flow control. In our paper, we consider a more complex business-to-business environment where the firm under consideration decides at each instant what price to set for a customer order. Further, we examine dynamic pricing in an environment of customer heterogeneity (core customers vs. fill-in customers).

Mendelson (1985) considers users of computing resources who have a positive delay cost and explicitly accounts for an externality cost (the delay cost that a user inflicts on all other users) and examines the optimal capacity choice. However, he considers only one static price to keep the queue length regulated. In a similar static setting, Mendelson & Whang (1990) consider internal pricing, i.e., transfer pricing for computer resources, and they address both resource

pricing (a price that determines if customers join the system), and priority pricing (set prices based on priorities). Results show that long jobs should be charged more than in proportion to their resource requirements. In their comments on future research, the authors state "it would also be of interest to study a dynamic version of our model, where current queue information is available and is being utilized in setting the resource and priority prices." This is, in part, what we do in this paper. Similarly, Dewan & Mendelson (1990) account for both a delay cost and a capacity cost where users have nonlinear delay cost functions. Customers and service requests are homogeneous. Users also decide to enter the system based on steady state queue lengths rather than on the current queue status, and the price is simply the opportunity cost of servicing that request. Stidham (1992) evaluates a variant of the Dewan and Mendelson model but allows for an upper bound on arrival rate. Customers cannot observe the state of the system on arrival, but a price is charged for admission. So the system designer must choose a service rate and a price (or arrival rate). The new constraint leads to a system in which the first order conditions may not lead to a globally optimal solution (see also Rump & Stidham (1998) and Kim & Mannino (2003)). Plambeck (2000), on the other hand, considers two classes of customers – delay sensitive and delay insensitive. Static prices are set for each class, but leadtimes are dynamically quoted. The heuristic policy sets a price and leadtime quote that induces heavy traffic. In this paper, we examine both static and dynamic pricing policies and evaluate the optimal solution analytically and consider a different segmentation of core versus fill-in customer segments.

In the marketing literature, many papers have appeared on dynamic pricing (Baker, Marn, & Zawada (2001), Hall, Kopalle, & Krishna (2002), Kopalle, Rao, & Assunção (1996), Kopalle, Mela, & Marsh (1999)), but most of that research has been in the area of consumer marketing (Gijsbrechts (1993)) and less so in the area of business-to-business marketing (Noble & Gruca (1999)). More importantly, the pricing research in marketing has typically ignored the supply chain implications – for example, Raman & Chatterjee (1995) consider the impact of demand uncertainty on the optimal price path without incorporating the corresponding upstream supply chain implications.

Thus, we propose an analytical, profit maximizing approach for a manufacturer's pricing decisions over time by considering the supply chain impact of dynamic pricing decisions. The proposed model considers customer heterogeneity, stochastic demand, and supply chain

(manufacturing capacity) constraints. The decision variables consist of a set of prices offered at different points in time that may be a function of the state of the manufacturing system. The contributions of this paper lie in (1) determining optimal dynamic pricing policies by taking into consideration the key supply chain issue of current congestion levels, and (2) characterizing the conditions under which one policy is superior to another via numerical simulations.

3. THE MODELS

3.1. Modeling Framework

We assume that a firm operating a production system faces two types of demand. The first is demand from "long-term" customers (herein also referred to as core customers) who enter into long-term agreements with the firm by specifying the average quantity required over time at a pre-established price. The second type of customer is a "short-term" / "brand switching" customer (herein also referred to as fill-in customers). These short-term customers approach the firm with immediate needs and will do business with the firm if they find the price and any production delays acceptable. (Seifert, Thonemann, & Hausman (2004) show that a firm can increase profits by procuring a moderate fraction of their demand on the spot market, i.e. as fill-in customers to the firm we model.) The production system is operated on a "make-to-order" basis, meaning that the firm does not hold inventory of finished or partly finished goods in advance of customer demand. Such a policy could follow from the economics of the situation (e.g., as in the MetalFab case with expensive finished goods and therefore risky inventory investments) or the characteristics of the products (customized on an order-by-order basis). We assume that both types of demand follow a Poisson arrival process. We denote the average arrival rate of core customers by λ_c and the average arrival rate of fill-in customers by λ_f. Given the long-term, contractual agreements between a firm and its core customers, we consider the arrival rates of core customers as given. However, the arrival rate for fill-in customers depends on the prices charged via a downward sloping demand function, $\lambda_f(p_f)$, where p_f is the price charged for fill-in customers. We assume that the time required to complete either core or fill-in orders is exponentially distributed with a mean of $1/\mu$.

We model the above production system as single-server queueing system. Although this framework is a simplification of the serial and parallel arrangement of workstations that exists in most production systems, our abstraction captures the qualitative nature of the congestion and delay that exist in real production systems in an analytically tractable manner. We assume that all jobs are handled on a first-come-first-served basis.

In addition to the revenue collected from customers, we assume that the firm is also concerned with the throughput time of customer orders. We capture this concern via a constraint on the expected throughput time of core customer orders, which we denote by W_c. We focus on waiting time for the core customers because these customers likely have entered into long-term agreements with expectations of a certain level of service or with a level of service contractually specified. Our use of expected throughput as a performance measure differs from typical use of a fixed, contracted delivery time. Alternative formulations might use a fractile of the throughput time distribution as a performance measure, e.g., 90% of jobs must be completed in 15 days or less. Recognizing that throughput time distributions can be difficult to characterize for many types of queueing systems, and recognizing that the first moment of a distribution captures much relevant information, we use the expectation. This approach may also be justified in a setting where customers are risk neutral and bear a linear delay cost (see, for example, Mendelson (1985)). In such a setting, the constraint on expected waiting time is also a constraint on the expected delay cost incurred. In other words, we have $W_c \leq W_0$, where W_c is the expected wait in the system for core customers and W_0 represents the maximum allowed expected wait for core customers. As in many dynamic settings (Kopalle et al. (1996)), the firm is interested in maximizing the sum of expected revenue over time. Using a per-period profit criterion (Bertsekas (1987)), this translates to maximizing average expected revenue collected per unit time. Below we present four models that consider different pricing policies for fill-in customers. They are:

1. State-independent pricing, which results in a constant price,
2. State-dependent pricing where jobs are admitted only when the system is idle,
3. State-dependent pricing where jobs are accepted up to a certain state and are charged a uniform price, and
4. General state-dependent pricing, in which no constraints are placed on the pricing policy.

Models 1 through 3 consider policies in which constraints are placed on the extent to which state information is utilized in making pricing decisions. These policies were chosen because they represent natural ways in which limited information might be used in practice.

3.2. Model 1: State-Independent Pricing

Here we study the case of a firm that sets a uniform price for fill-in jobs independent of the current state of the factory. We do so for benchmarking purposes because such a static policy is simple and requires no real-time information about factory status. It does, however, require information about time-average system behavior, as might be available from historical data. This policy serves as a baseline to compare the performance of our other policies that make use of system state information. State-independent pricing is also consistent with the behavior of core customers whose arrival rate is independent of the system state. This is based on the nature of long-term contracts that often specify a fixed price and a service level commitment, where the supplier manages its production system to satisfy those commitments.

In this context, the problem faced by the supplier is to maximize the expected revenue collected from fill-in customers per unit time, i.e.,

$$\max_{p}\{\lambda_f(p)p\} \tag{1}$$

subject to:

$$W_c \le W_0 .$$

The simple state independent problem outlined above can be modeled as an M/M/1 queue with arrival rate given by $\lambda_c + \lambda_f(p)$ and service rate given by μ. Accordingly, we have (see, e.g., Gross & Harris (1985)):

$$W_c = \frac{1}{\mu - (\lambda_c + \lambda_f(p))} \tag{2}$$

We assume $\mu - \lambda_c > 0$ and $\frac{1}{\mu - \lambda_c} \le W_0$, which implies that the expected wait in the absence of fill-in arrivals is less than the upper bound on expected wait, which is necessary for feasibility. Note that W_c and W_f (where W_f is the expected throughput time for fill-in customer orders) are equal for this policy. Problem (1) can be solved via Lagrangean methods (where β denotes the

Langrage multiplier). The solution consists of two possible cases depending on whether the waiting time constraint is binding. The solution is presented below as Theorem 1.

Theorem 1. State-Independent Pricing.

(a) For problem (1), we have the following two cases:

Case 1: $\beta = 0$ and the optimal price p^* satisfies $\lambda_f'(p^*)p^* + \lambda_f(p^*) = 0$;

Case 2: $\beta = \left(\frac{1}{(\mu - \lambda_c - \lambda_f(p^*))}\right)^2 \left(p^* + \frac{\lambda_f(p^*)}{\lambda_f'(p^*)}\right)$ and the optimal price p^* satisfies

$\lambda_f(p^*) = \mu - \lambda_c - \frac{1}{W_0}$.

(b) Under linear demand, the first-order conditions are sufficient for optimality.

All proofs are in the Appendix.

Cases 1 and 2 in Theorem 1 correspond to the instances where the waiting time constraint is either not binding or binding, respectively. If the waiting time constraint is not binding, the solution is simply that of a standard monopolist in which all the operational complexities of the system can be ignored, i.e., in which the firm is interested only in maximizing $p\lambda(p)$. This solution is simple and it reveals to us what a job is worth. The term $\lambda_f'(p^*)p^*$ measures the cost of increasing price by accounting for the value of lost demand that results from a price increase. The factor $\lambda_f'(p^*)$ is the reduction in demand for a unit increase in price, and the factor p^* is the return earned per unit of demand. Here, the value of a job is simply p^*, but we will demonstrate that this factor takes on more complex forms which yield interpretations that can help managers understand both the benefits and costs of taking on additional work.

If the waiting time constraint is binding, the Lagrange multiplier β provides a measure of the cost of the constraint. This cost enters into the calculation of the optimal price in a way that assesses an opportunity cost for each job that is accepted based on its impact on the expected waiting time. In other words, we have,

$$\lambda_f(p^*) + \lambda_f'(p^*)\left(p^* - \beta\frac{\partial W_c}{\partial \lambda_f}\right) = 0\,. \tag{3}$$

Note that if there were a variable cost c associated with performing each job and operational complexities were ignored, then the first-order condition for the optimal price would be identical to (3) with c replacing $\beta\frac{\partial W_c}{\partial \lambda_f}$. This is consistent with our interpretation of $\beta\frac{\partial W_c}{\partial \lambda_f}$ as a variable cost associated with each job. A numerical example is presented below.

Example 1. Assume fill-in customers exhibit a linear demand curve of the form $\lambda_f(p) = 100 - 0.1p$. Jobs from core customers arrive at an average rate of 8 per month and the production system can complete 10 jobs per month on average if continuously busy. The firm imposes an expected waiting time constraint of one month for core customers.

Ignoring the operational complexities of this problem would result in choosing a price of $500 per fill-in arrival, resulting in 50 such jobs per month on average. Clearly such a scheme would overwhelm the firm's productive capacity of 10 jobs per month, and we can conclude that the waiting time constraint must be binding. Using the results of Case 2 from Theorem 1, we conclude that the optimal arrival rate is a single fill-in job per month on average at a price of $990. Even though the waiting time constraint is binding, the firm should still accept the occasional, high revenue job. The value of the Lagrange multiplier β is here $980.

3.3. Model 2: State-Dependent Pricing – Admitting Jobs When Idle

In this subsection we study, as a first-step towards developing more complex dynamic pricing policies, a simple form of dynamic pricing where fill-in arrivals are allowed only when the system is otherwise idle. This policy captures the idea that a factory manager may wish to only accept fill-in work when the factory is relatively non-congested. Under this policy, the system under study is a Markov chain with departure rate μ in every state and arrival rate λ_c in every state except state zero, where the arrival rate is $\lambda_c + \lambda_f$. Lemma 1 below summarizes some basic properties of such a system.

Lemma 1. Admit Fill-in Jobs When Idle – Properties. For the Markov chain described above, we have:

(a) The steady-state probability that the system is idle, denoted by π_0 is given by:

$$\pi_0 = \frac{\mu - \lambda_c}{\mu + \lambda_f};$$

(b) The expected waiting time in the system for core customer arrivals is given by:

$$W_c = \frac{\pi_0}{\mu}\left[1 + \left(\frac{\lambda_c + \lambda_f}{\mu}\right)\left(\frac{1}{\left(1 - \frac{\lambda_c}{\mu}\right)^2} + \frac{1}{\left(1 - \frac{\lambda_c}{\mu}\right)}\right)\right].$$

For this policy, the expected waiting time for fill-in customers is just their service time, since they only enter the system when it is idle. The problem faced by the firm in this case is to maximize the expected revenue from fill-in customers per unit time:

$$\max_p \{\pi_0 \lambda_f(p) p\} \tag{4}$$

subject to:

$$W_c - W_0 \le 0.$$

Note that the firm potentially collects revenue from fill-in customers only when the production system is idle, as captured by the factor π_0 in (4). Theorem 2 below summarizes the solution for problem (3).

Theorem 2. Admit Fill-In Jobs When Idle – Solution.

(a) For problem (4), we have the following two cases:

Case 1: $\beta = 0$ and the optimal price p^* satisfies $\lambda_f'(p^*)\left(p^* - \frac{\Pi}{\mu - \lambda_c}\right) + \lambda_f(p^*) = 0$, where

$\Pi = \pi_0 \lambda_f(p^*) p^*$ represents the optimal expected return per unit time from fill-in jobs;

Case 2: $\beta \neq 0$ and the optimal price p^* and β satisfy

$$\lambda_f'(p^*)\left(p^* - \frac{\Pi}{\mu - \lambda_c} - \frac{\beta}{\pi_0}\frac{\partial W_c}{\partial \lambda_f}\right) + \lambda_f(p^*) = 0 \text{ and}$$

$$W_0 = \frac{\mu - \lambda_c}{\mu\left(\mu + \lambda_f\left(p^*\right)\right)}\left[1 + \left(\frac{\lambda_c + \lambda_f\left(p^*\right)}{\mu}\right)\left(\frac{1}{\left(1 - \frac{\lambda_c}{\mu}\right)^2} + \frac{1}{\left(1 - \frac{\lambda_c}{\mu}\right)}\right)\right].$$

(b) For linear demand, the first-order conditions are sufficient for optimality.

Note that for this form of state-dependent pricing, an opportunity cost enters into the calculation of p^* in Theorem 1 even when the waiting time constraint is non-binding. This opportunity cost, which also appears in the expression for p^* when the waiting time constraint is binding, takes the form:

$$\frac{\Pi}{\mu - \lambda_c}, \tag{5}$$

where Π is the optimal return from the fill-in jobs per unit time. We note that:

$$\frac{\partial\left(\frac{1-\pi_0}{\pi_0}\right)}{\partial \lambda_f} = \frac{1}{\mu - \lambda_c}. \tag{6}$$

The above factor measures the change in the relative odds that the system is in the set of states in which fill-in arrivals are not allowed. Thus, (6) measures the expected lost revenue that results from increasing the arrival rate and its subsequent effect on the steady-state probabilities.

For the case where the waiting time constraint is binding, a second opportunity cost, $\frac{\beta}{\pi_0}\frac{\partial W_c}{\partial \lambda_f}$, appears in the first-order condition. This cost can be interpreted as capturing the waiting time externality caused by a job arrival. The factor, $\beta\frac{\partial W_c}{\partial \lambda_f}$, captures the cost of the increased expected waiting time for core arrivals (via the shadow price of the waiting time constraint β) due to an increase in λ_f of one unit. This cost is translated into a cost that is relevant for an actual arrival via the factor $\frac{1}{\pi_0}$, which measures the expected number of fill-in arrivals until one finds the system idle and is admitted. A numerical example is presented below.

Example 2. Again, we assume the potential short-term customers exhibit demand of the form $\lambda_f(p) = 100 - 0.1p$; core customers arrive at an average rate of 8 per month and the production system can complete 10 jobs per month on average. The firm imposes an expected waiting time constraint for core customers of one month.

If the waiting time constraint is non-binding, the first-order condition for p^* is a quadratic equation with only one feasible solution: $p^* \cong \$768.33$. This price results in an arrival rate for fill-in jobs of 23.17 jobs per month. The resultant waiting time is found to be approximately 0.57 months, which leads to a non-binding waiting time constraint, as initially assumed. The probability that this system is idle is approximately 0.0603, and thus the expected additional revenue due to the fill-in jobs is approximately \$1073 per month. It is interesting to note that the expected performance of this relatively simple form of state-dependent pricing exceeds that of state-independent pricing by approximately 8.4% in this example. (As discussed in Section 4 below, this policy does not always outperform simple state-independent pricing.) The chief benefit of this policy is its relative simplicity: managers can easily determine if the factory is idle and communicate that fact to marketing personnel who are evaluating potential fill-in customers.

3.4. Model 3: State-Dependent Pricing – Constant Price Up To Cutoff State

Here we study a pricing scheme in which fill-in job arrivals are allowed only if the system is in a relatively uncongested set of states and all fill-in arrivals are charged a uniform price. Such a scheme is simpler to implement in practice than general state-dependent pricing (discussed in the next subsection), since it entails only a check to see if the factory is congested or not, e.g., by examining the number of outstanding orders. If the factory is too busy, the work is turned away. Otherwise, the work is accepted at a standard price for fill-in jobs. This policy generalizes Model 2 by endogenizing the cut-off state for admitting jobs. We denote the cutoff state as s, meaning that new fill-in work is accepted if there are s or fewer total jobs in the system. Note that the cutoff state, s, is a second decision variable, along with price, p. (We note that a simplified version of this model that relies upon only s as a decision variable could

prove useful in situations where the firm must operate as a price-taker.) Lemma 2 below presents the properties of such a queueing system.

Lemma 2. Acceptance Up To Cutoff State – Properties. For the Markov chain described above, we have:

(a) The steady-state probability that the system is in state s or lower, denoted by $F(s)$, is:

$$F(s)=\begin{cases}(s+1)\pi_0 \text{ if } \lambda_c+\lambda_f=\mu \\ \left[\dfrac{1-\left(\dfrac{\lambda_c+\lambda_f}{\mu}\right)^{s+1}}{1-\left(\dfrac{\lambda_c+\lambda_f}{\mu}\right)}\right]\pi_0 \text{ otherwise;}\end{cases}$$

(b) The expected waiting time in the system for core arrivals is given by:

$$W_c=\frac{\pi_0}{\mu}\left(\sum_{i=0}^{s+1}\frac{(i+1)(\lambda_c+\lambda_f)^i}{\mu^i}+\sum_{i=s+2}^{\infty}\frac{(i+1)(\lambda_c+\lambda_f)^{s+1}\lambda_c^{i-(s+1)}}{\mu^i}\right);$$

where

$$\pi_0=\begin{cases}\left(s+1+\dfrac{1}{1-(\lambda_c/\mu)}\right)^{-1} \text{ if } \lambda_c+\lambda_f=\mu \\ \left(\dfrac{1-\left(\dfrac{\lambda_c+\lambda_f}{\mu}\right)^{s+1}}{1-\left(\dfrac{\lambda_c+\lambda_f}{\mu}\right)}+\dfrac{\left(\dfrac{\lambda_c+\lambda_f}{\mu}\right)^{s+1}}{1-\left(\dfrac{\lambda_c}{\mu}\right)}\right)^{-1} \text{ otherwise.}\end{cases}$$

It is straight-forward to show that $W_f \le W_c$ here, so the waiting time constraint for core arrivals is also an upper bound on expected waiting time for fill-in jobs. The problem faced by the firm in this case is to maximize the revenue from fill-in customers:

$$\max_{p,s}\{F(s)\lambda_f(p)p\} \tag{7}$$

subject to:

$$W_c - W_0 \le 0\,.$$

Note that the firm potentially collects revenue from fill-in customers only when the production system is in states 0 through s, as captured by the factor $F(s)$ in (7). Theorem 3 summarizes the solution for problem (7).

Theorem 3. Acceptance Up To Cutoff – Solution.

(a) For problem (7), we have the following two cases:

Case 1: $\beta = 0$, s^* is infinite, and the optimal price p^* satisfies $\lambda_f'(p^*)(p^*) + \lambda_f(p^*) = 0$;

Case 2: $\beta \neq 0$ and the optimal price p^*, s^* and β satisfy:

$$\lambda_f'(p^*)\left(p^* - \Pi \frac{\partial\left(\frac{1-F(s)}{F(s)}\right)}{\partial \lambda_f} - \frac{\beta}{F(s)} \frac{\partial W_c}{\partial \lambda_f} \right) + \lambda_f(p^*) = 0, \text{ where } \Pi = F(s)\lambda_f(p^*)p^*;$$

s^* satisfies:

$$\frac{F(s^*+1) - F(s^*)}{W_c(s^*+1) - W_c(s^*)} \leq \frac{\beta}{\lambda_f(p)p} \text{ and } \frac{F(s^*) - F(s^*-1)}{W_c(s^*) - W_c(s^*-1)} \geq \frac{\beta}{\lambda_f(p)p};$$

and

$$W_c = W_0.$$

(b) For linear demand, there exists a unique optimal value of p for each s, denoted $p^*(s)$.

We have not demonstrated that problem (7) is jointly quasi-concave in p and s, although we have demonstrated that $p^*(s)$ is unique. The problem thus reduces to a one-dimensional search over possible positive integer values of s. Further, in all of our numerical studies, the reduced problem in $p^*(s)$ and s was found to be quasi-concave in s.

Of note in Theorem 3 is that for a non-binding waiting time constraint, it is optimal to use state-independent pricing for fill-in jobs. From a managerial standpoint, if the factory is very lightly loaded or customers are relatively insensitive to delays, it is best to ignore the state of the factory in deciding which jobs to accept – only price should be used to discourage or encourage customers.

Also of note are the two "opportunity costs" that arise in the first-order condition for the optimal price. Both of these costs are similar to the case where fill-in orders are admitted only when idle (see Theorem 2). Note that admitting fill-in jobs only when idle corresponds to $s = 0$ in this case. The term which incorporates the waiting time constraint, $\frac{\beta}{F(s)}\frac{\partial W_c}{\partial \lambda_f}$, has the same form and interpretation as for Theorem 2. The term $\Pi \frac{\partial((1-F(s))/F(s))}{\partial \lambda_f}$ is also similar in form to the result of Theorem 2, except that $F(s)$ replaces π_0. Here, Π again represents the optimal expected return per unit time from fill-in jobs. The expression $(1-F(s))/F(s)$ gives the relative odds that the production system is in a state in which no fill-in arrivals are accepted. The change in these relative odds multiplied by the optimal revenue from fill-in jobs provides a measure of the opportunity cost of accepting another job, i.e., it captures the long-term consequences of the decision. In other words, if the change in relative odds is high when the fill-in arrival rate increases by one unit, then the opportunity cost of accepting a job is high. This implies that the manager should charge a relatively higher price for fill-in jobs in such a setting. A numerical example is presented below.

Example 3. Again, we assume fill-in customers exhibit demand of the form $\lambda_f(p) = 100 - 0.1p$; core customers again arrive at an average rate of 8 per month and the production system can complete 10 jobs per month on average. The firm imposes an expected waiting time constraint for core customers of one month.

We know from Example 1 that $\beta \neq 0$, and thus the waiting time constraint must be binding. Optimizing using Excel™ Solver™ yields $s^* = 6$ (i.e., fill-in jobs are admitted when there are six or fewer jobs currently in the system) and $p^* \cong \$936.82$, which equates to $\lambda_f \cong 6.32$ fill-in jobs per month. The optimal expected profit per month from fill-in jobs is approximately \$1767.

Note that, for the above example, this policy yields a significant improvement in expected profitability over the policy of admitting fill-in jobs only when idle – the gain is approximately 65%! (The gain is 78.5% in comparison to the state-independent pricing policy.)

However, this policy is slightly more complicated to implement in practice, for it requires knowledge of the factory state, beyond simply knowing whether the factory is idle. But the gains are remarkable. The typical request from Marketing is to take every job, whereas the response from Manufacturing is often "we are too busy." This example helps determine what "busy" means, and it illustrates the potential gains from optimally choosing the cut-off state and the price jointly.

3.5. Model 4: General State-Dependent Pricing

In this subsection we study a general state-dependent pricing scheme in which price may be changed dynamically without constraint, i.e., fill-in job arrivals at time *t* are charged a price that is a function of the congestion levels in time *t*. Of course, such a scheme entails tracking the status of the production process carefully so that accurate prices can be specified. We use λ_{fi} to represent the fill-in arrival rate when there are i customers in the system. Lemma 3 below presents the properties for such a queueing system.

Lemma 3. General State-Dependent Pricing – Properties. For the Markov chain described above, we have:

(a) The steady-state probability that the system is in state i, denoted by π_i, is given by:

$$\pi_i = \frac{\prod_{j=0}^{i-1}\left(\lambda_c + \lambda_{fj}\right)}{\mu^i}\pi_0,$$

where

$$\pi_0 = \left[1 + \sum_{i=1}^{\infty}\left(\frac{\prod_{j=0}^{i-1}\left(\lambda_c + \lambda_{fj}\right)}{\mu^i}\right)\right]^{-1};$$

(b) The expected waiting time in the system for core arrivals is given by:

$$W_c = \frac{\pi_0}{\mu}\left[1 + \sum_{i=1}^{\infty}\left((i+1)\frac{\prod_{j=0}^{i-1}\left(\lambda_c + \lambda_{fj}\right)}{\mu^i}\right)\right].$$

The objective of the firm in this case is to maximize the expected revenue per unit time from fill-in customers:

$$\max_{p_0, p_1, p_2, \ldots} \left\{ \sum_{j=0}^{\infty} \pi_j \lambda_{fj}(p_j) p_j \right\} \tag{8}$$

subject to:

$$W_c - W_0 \leq 0 .$$

Note that in this setting we are searching for an infinite-dimensional vector of optimal prices. Theorem 4 below summarizes the solution for problem (8).

Theorem 4. General State-Dependent Pricing – Solution.

(a) For problem (8), we have the following two cases:

Case 1: $\beta = 0$ and the optimal prices p_j^* satisfy $\lambda'_{fj}(p_j^*)(p_j^*) + \lambda_{fj}(p_j^*) = 0$ for all j;

Case 2: $\beta \neq 0$ and the optimal prices p_i^* satisfy:

$$\lambda'_{fi}(p_i^*)\Bigg[p_i^* - \left(\Pi - \lambda_{f,i+1}(p_{i+1}) p_{i+1} + \beta\left(\frac{i+2}{\mu} - W_c \right) \right) \frac{\partial(\pi_{i+1}/\pi_i)}{\partial \lambda_{fi}} - \left(\Pi - \lambda_{f,i+2}(p_{i+2}) p_{i+2} + \beta\left(\frac{i+3}{\mu} - W_c \right) \right) \frac{\partial(\pi_{i+2}/\pi_i)}{\partial \lambda_{fi}} - \left(\Pi - \lambda_{f,i+3}(p_{i+3}) p_{i+3} + \beta\left(\frac{i+4}{\mu} - W_c \right) \right) \frac{\partial(\pi_{i+3}/\pi_i)}{\partial \lambda_{fi}} - \ \ldots \ \Bigg] + \lambda_{fi}(p_i^*) = 0,$$

for all i, where

$$\Pi = \sum_{j=0}^{\infty} \pi_j \lambda_{fj}(p_j^*) p_j^* ;$$

and

$$W_c = W_0 .$$

(b) For linear demand, the first-order conditions are sufficient for optimality for $\beta = 0$.

Given the complexity of this problem, we have not established conditions under which the first-order conditions are sufficient for optimality when $\beta \neq 0$. For the case $\beta \neq 0$, it is simple to show that only a finite number of states will exhibit a non-zero arrival rate under linear demand,

which simplifies the search for a solution. Under these same conditions, the prices (and arrival rates) will be monotone in the state (see Theorem 5 below). In our experience with numerical examples, we found no instances in which local optima were not global.

Of note in Theorem 4 is that, just as for Theorem 3, if the waiting time constraint is non-binding, it is optimal to use a single state-independent price for all fill-in jobs. In other words, if customers are not sensitive to delays or if there is significant unused capacity in the factory, a "set and forget" approach is optimal, i.e., a single uniform price is set and factory status is ignored. As a result, implementation is much simpler for lightly loaded systems than for heavily loaded systems.

For the case where the waiting time constraint is binding, we have an infinite series of terms included within the first order conditions that represent the "costs" of an additional arrival. Not surprisingly, these costs bear a resemblance to similar costs that appear in Theorems 2 and 3. Each term in this infinite series tabulates the impact of an additional arrival in the state under consideration upon all "larger" states. The first of these terms is:

$$\left(\Pi - \lambda_{f,i+1}(p_{i+1})p_{i+1} + \beta\left(\frac{i+2}{\mu} - W_c\right)\right)\frac{\partial(\pi_{i+1}/\pi_i)}{\partial\lambda_{fi}} \tag{9}$$

The expression $\Pi - \lambda_{f,i+1}(p_{i+1})p_{i+1}$ measures the difference between the optimal revenue per unit time and the revenue per unit time when in state $i+1$. The term

$$\beta\left(\frac{i+2}{\mu} - W_c\right) \tag{10}$$

measures the difference between the expected waiting time in the system for a core customer arrival when the system is in state $i+1$ (note that this is just the expected time to service the existing $i+1$ customers plus the newly arrived customer) and the overall expected wait in the system for core customers, given by W_c. The Lagrange multiplier β is multiplied by this difference to yield a "relative" cost of the waiting time constraint when in state $i+1$. Note that this cost can be negative when i is small. The factor

$$\frac{\partial(\pi_{i+1}/\pi_i)}{\partial\lambda_{fi}} \tag{11}$$

represents the change in the relative odds of occupying state $i+1$ due to an increase in the arrival rate in state i. Similar interpretations can be applied to subsequent terms in the series.

Theorem 5 below demonstrates that the set of optimal prices for problem (8) are monotone in the state.

Theorem 5. General State-Dependent Pricing – Monotonicity. Under linear demand, the solutions to problem (8) satisfy $p_i^* \le p_{i+1}^*$ for all i.

This monotonicity result supports our intuition; as a production system becomes more congested, a higher price should be charged for fill-in jobs because of the greater strain that they put on the system. Using the result of Theorem 5, it is straightforward to show that $W_f \le W_c$, so as with the previous policies considered, the waiting time constraint for core arrivals is also an upper bound on expected waiting time for fill-in jobs.

A numerical example illustrating this policy is presented below. The basic setting is the same as prior examples.

Example 4. Following Examples 1-3, fill-in customers exhibit demand of the form $\lambda_f(p) = 100 - 0.1p$; core customers arrive at an average rate of 8 per month and the production system can complete 10 jobs per month on average. The firm imposes an expected waiting time constraint of one month for all customers.

We know from Example 1 that $\beta \neq 0$, and thus the waiting time constraint must be binding. Optimization using Excel™ Solver™ yields the results shown in Table 1 below. The progression of prices and arrival rates shown in Table 1 confirms our intuition about these relationships, namely that as the system becomes more congested, a higher price is charged for fill-in arrivals and demand is curtailed. The optimal expected profit per month from fill-in jobs is approximately $1840. It is notable that the improvement relative to the "uniform price up to a cut-off point" policy (from Example 3) is only about 4.1%. This gain comes at the cost of significant additional monitoring of factory status required in order to inform the price choice for this general state-dependent pricing scheme. Without sophisticated tracking software, most managers do not have access to the information required to implement this scheme. It is thus encouraging that the incremental gain is quite small.

INSERT TABLE 1 ABOUT HERE

4. POLICY COMPARISON

Figure 1 provides a comparison of the optimal fill-in job arrival rates under each of the four examples discussed in Section 3, plotted with respect to system state, i.e., the number of jobs in queue plus any being served. Clearly the state-independent or static policy yields a conservative approach to fill-in customers; a relatively low arrival rate is maintained to compensate for lack of state information. The policy of admitting fill-in work only when idle yields a contrasting approach. In this example, the waiting time constraint on core jobs is non-binding, so the constraint of interest is that created by the choice of policy. The result is that a relatively high arrival rate is maintained in the sole state in which we are allowing jobs to be admitted in order that this condition can be exploited. The policy of admitting at a uniform price up to a cut-off state yields, in this example, an intermediate level of arrivals in those states where arrivals are allowed. Lastly, the uniform state-dependent policy yields a set of arrival rates which bracket those of the prior policy, taking better advantage of opportunities when the system is relatively idle and better recognizing the costs of additional arrivals when the system is relatively congested. While the "uniform up to a cut-off" and "general state dependent pricing" policies look quite different from this perspective, it is notable that their relative financial performance is much less divergent. This result was discussed in Section 3 for this particular numerical example and is explored in more depth in this section.

INSERT FIGURE 1 ABOUT HERE

In the remainder of this section, we expand our set of numerical examples to study policy performance under a broad range of conditions. Expanding on Examples 1-4, we vary two basic model characteristics: (1) the system utilization due to core (long-term) customers, denoted by ρ_c, where $\rho_c = \lambda_c / \mu$, which we vary between 0 and 0.9 (0.9 is the point at which the waiting time constraint is binding and no additional arrivals are feasible); and (2) the form of the demand function for fill-in customers, which takes on two forms representing either a relatively "large"

market or a relatively "small" market. The large market demand function has the same form as in the Section 3 examples: $\lambda_f(p) = 100 - 0.1p$. The small market demand function has one tenth the magnitude: $\lambda_f(p) = 10 - 0.01p$.

Figure 2 is a plot of the optimal expected profit earned from fill-in jobs as a function of the system utilization due to core jobs. These solutions correspond to the "large" market demand function. The revenue-maximizing demand for this demand function, ignoring the operational details of the problem, is $\lambda_f^* = 50$ jobs per month. This would clearly overwhelm the productive capacity of the firm in these examples (recall $\mu = 10$ jobs per month), and explains why we refer to this as a "large" market demand function. As mentioned above, when core job utilization is 0.9, the waiting time constraint is binding and fill-in arrivals are not feasible, as revealed in Figure 2. Figure 2 also shows that general state-dependent pricing sets a performance frontier that encompasses the other policies. Of course, the feasible set of general state-dependent pricing solutions is a superset of the feasible solutions under each of the other policies, and therefore the solution must be at least as good as any of the others.

An interesting aspect of Figure 2 is the relatively poor performance of the policy of admitting jobs only when the system is idle. At $\rho_c = 0.8$, this policy outperforms state-independent pricing by 8.3%, but for all lower plotted levels of ρ_c, this policy is dominated by state-independent pricing. The explanation for this effect is that admitting jobs only when idle gets costly for a lightly loaded system; such a policy does not allow any queueing of fill-in jobs, which can result in unnecessary idle time. Conversely, for heavily loaded systems, admitting jobs only when idle does relatively well compared to a state-independent policy because in such a case it becomes important to account for externality effects, i.e., the impact of an accepted fill-in job on the delay of the core jobs.

INSERT FIGURE 2 ABOUT HERE

Figure 2 also reveals that the relative value of using internal system state information to make job pricing and acceptance decisions increases as ρ_c increases. Figure 3 provides more detail by presenting a plot of the relative gain that results from a full state-dependent pricing

policy vs. a state-independent policy for a range of values of ρ_c. For $\rho_c = 0$, the relative gain from a full state-dependent policy is only 8.6%, whereas for $\rho_c = 0.895$, the relative gain is 812%. Of course, as can be seen in Figure 2, the absolute return earned under either policy declines as ρ_c increases. The large relative gain results because the performance gap between the policies remains significant even as the absolute gains decline.

While the relative gains from full state-dependent pricing can be tremendous, the costs of using a policy of uniform pricing up to a cut-off state instead of full state-dependent pricing are much more moderate. Figure 4 plots the relative gain of uniform pricing up to a cut-off vs. a state-independent policy. The curve is very similar to that of Figure 3 for general state-dependent pricing. Among the solutions plotted in Figures 3 and 4, the maximum performance improvement that general state-dependent pricing yields relative to uniform pricing up to a cut-off point is 13.8% at $\rho_c = 0.89$, and the gain is only 3-6% over most of the range of ρ_c (the difference in relative performance is not monotone in ρ_c, which can be explained by the integer constraint imposed on the cut-off state s). These gains are quite modest when compared to the value of using internal state information, i.e., the gains from implementing a state-dependent pricing policy vs. a state-independent policy.

INSERT FIGURE 3 ABOUT HERE

INSERT FIGURE 4 ABOUT HERE

Figure 5 illustrates the results of a numerical analysis that is identical to that plotted in Figure 2, except that the demand function for fill-in arrivals has the form $\lambda_f(p) = 10 - 0.01p$. We refer to this as a "small" market demand function because the revenue maximizing arrival rate, ignoring all operational details, is $\lambda_f^* = 5$, which is a feasible arrival rate with $\mu = 10$. In fact, this is the solution depicted in Figure 5 for three of the policies when $\rho_c \leq 0.4$. For these cases, the waiting time constraint is non-binding and the optimal solution for the policies of uniform pricing up to a cut-off state and general state-dependent pricing both collapse to the state-independent pricing solution, as described in Theorems 3 and 4. As ρ_c increases, the

waiting time constraint becomes binding and these two state-dependent policies begin to differ from and outperform state-independent pricing.

INSERT FIGURE 5 ABOUT HERE

Figure 6 presents the relative gains from general state dependent pricing vs. state-independent pricing, and Figure 7 presents the relative gains from uniform pricing up to a cut-off state vs. state-independent pricing. These plots are analogues of Figures 3 and 4 for the small market case. Figure 6 reveals that the relative value of state information increases dramatically as the system becomes more congested. Figure 7 again supports the idea that the relative gains from full-state dependent pricing vs. uniform pricing up to a cutoff are limited – here the maximum gain is 6.9% at $\rho_c = 0.88$.

INSERT FIGURE 6 ABOUT HERE
INSERT FIGURE 7 ABOUT HERE

Although our numerical analysis cannot serve as proof of any normative statements regarding these four policies, they do provide some important insights. In all cases, a policy of uniform pricing up to a cut-off state performs quite well versus full state-dependent pricing. This is particularly noteworthy for a number of reasons: (1) it requires less detailed information about the state of the system, i.e., management only needs to know whether the factory is "too busy to accept fill-in work" or not; (2) the uniform pricing policy is much easier to compute and communicate to relevant managers; and (3) a uniform pricing policy is less likely to cause dissonance among customers who experience widely varying prices that result from a process over which they have no control. Another observation that follows from our analysis is that the relative value of using information about factory status increases dramatically as the factory becomes more congested. This result is not surprising, but the relative magnitude of the effect can be huge, ranging up to 812% as noted above.

An analytic result that lends some support to the numerical results of this section is provided via the concept of entropy from the field of information theory. Entropy is a measure of uncertainty in a probabilistic system, i.e., it is a measure of the complexity required to

describe a random variable, and serves as an indicator of the minimum complexity of the system required to produce instances of a given random variable. (See, for example, Cover and Thomas (1991) for a discussion of entropy). Entropy is formally defined for a positive, discrete random variable as:

$$H = -\sum_{i=0}^{\infty} p_i \log_2(p_i),$$

where the result is measured in "bits" when the logarithm is taken as base 2, as here. It is well known that the steady-state state distribution for the M/M/1 queue follows a geometric distribution (see, for example, Gross & Harris (1985)). This result can be used to calculate the entropy associated with the state distribution for an M/M/1 queue:

$$H = -\log_2(1-\rho) - \frac{\rho}{1-\rho}\log_2(\rho),$$

where ρ is the system utilization. When $\rho = 0$, we have $H = 0$ bits (we define $0\log(0) \equiv 0$). In other words, if the system is always idle, there is no uncertainty in its state. When $\rho = 0.5$, we have $H = 2$ bits and when $\rho = 0.85$, we have $H \cong 4$ bits. Further, it can be shown that $\lim_{\rho \to 1} H = \infty$ for the M/M/1 queue. Figure 8 plots the system state entropy as a function of system utilization. Clearly, system entropy increases in system utilization. Therefore, more information is revealed from observing the state of a highly utilized system than would be revealed by observing a less highly utilized production system. Hence, there is substantive information revealed when observing the system state when the utilization is high. This is precisely the reason why we observe increasing relative gains with state-dependent pricing versus a state-independent pricing as core customer utilization increases (see Figures 3, 4, 6 and 7). The shape of the curve in Figure 8 closely resembles that of Figures 3, 4, 6, and 7, all of which measure the impact of state knowledge in terms of relative financial return (and which reflect queueing systems that depart somewhat from the assumptions of an M/M/1 system). Figure 8 reveals that there is an underlying concept of the information content of state information that closely parallels the value of that information revealed in our numerical examples. This analytic result also provides some assurance that the results of our set of numerical examples are somewhat general in nature.

INSERT FIGURE 8 ABOUT HERE

The interpretation of entropy as a measure of the complexity of the system required to generate a random variable can also provide some insights into the performance of the four pricing policies considered in this paper (see, for example, Cover and Thomas (1991) for an explanation of Kolmogorov complexity). Each of these policies (except static pricing as in Model 1) requires a signal from the factory floor in order to establish the current price for fill-in work. The entropy associated with these signals is a measure of the minimum complexity of the systems required to generate these signals, which should correlate with the relative difficulty of implementing such a system on the factory floor.

We can revisit the examples of Section 3 and use this notion of complexity to evaluate the performance/complexity tradeoff for the policies we consider. In the examples of Section 3, state-independent pricing requires no information about factory status to quote a price. Admitting only when idle requires 0.329 bits of information per factory status inquiry on average, uniform price up to a cutoff state requires 0.880 bits, and general state-dependent pricing requires 4.172 bits. (These figures can be calculated from the state probabilities that are relevant to the decision; in the case of admitting only when idle, we need only be concerned with the states "idle" and "busy," which are occupied with probabilities 0.060 and 0.940, respectively.) These entropy values can be used to calculate a "return per bit" from each of the three dynamic policies. Table 2 summarizes the results of this calculation for the examples of Section 3. Here again, we see evidence that the policy of uniform pricing up to a cutoff state deserves attention: it far outperforms the other policies on this metric. We've demonstrated previously that this policy performs nearly as well as general state-dependent pricing in a variety of numerical studies. Here see that this performance is coupled with a relatively low measure of policy complexity, indicating that policy is likely to be relatively easy to implement.

INSERT TABLE 2 ABOUT HERE

5. CONCLUSIONS AND FUTURE RESEARCH

The objectives of this paper are twofold. First, although there has been much research in the area of dynamic pricing, very few papers have integrated supply chain issues with pricing policies. This paper contributes to the literature that attempts to fill this gap. Second, the papers that do integrate supply chain issues with pricing policies generally focus on inventory levels. Little is known, however, with respect to how pricing policies and the supply chain dynamically interact in make-to-order scenarios. Our work sheds some light on these issues. In particular, we combine the study of two types of customers in a natural way. Long-term, or core, customers face state-independent pricing and a constraint on expected order delay, consistent with typical long-term contractual arrangements. Much more flexibility is generally allowed in choosing pricing policies for short-term, or fill-in, customers, and this choice forms the core of this paper.

We introduced four models of the interaction of pricing and the factory status for fill-in customers, including (1) ignoring queue length and setting a single, constant price for fill-ins, (2) admitting fill-ins only when the system is idle and setting a price accordingly, (3) admitting fill-ins up to a system state of s under a uniform price, and (4) setting a different price for each possible state of the system. Models 1 through 3 considered policies in which constraints are placed on the extent to which state information is utilized in making pricing decisions. We find the optimal prices analytically for each model, and for model (3) we also find the optimal cut-off state, s^*. We also analytically illustrate insights and properties common to the four models, taking into consideration (a) the waiting time constraint for the core customers and (b) the opportunity cost of taking up an additional fill-in order. Finally, we illustrate the behavior of the models using numerical analyses, which highlight new insights.

Clearly, the policy that requires the least information from production is state-independent pricing. One would expect, therefore, that any policy which employs some information about queue status would show higher profitability. This is not necessarily true, however. For instance, a policy that admits fill-ins only when the system is idle can perform much worse than state-independent pricing when system utilization is relatively low. In other words, using only information about the factory being idle can actually be worse than using no factory status information at all.

Perhaps the most striking, and encouraging, result is that a policy that admits fill-in customers at a constant price up to a cut-off state of s, where both s and price are set optimally, can dramatically outperform a state-independent policy – yielding an almost 80% gain in the

example of Section 3. Using queue information in this case is likely well worth the effort. Further, the increase in profit achievable from setting a different optimal price for each possible state of the production system, relative to the policy of a uniform price up to a cut-off state, is fairly small – about 4% in the same example cited above. This result suggests that while managers can improve profit markedly by using information about the status of the factory, a fairly simple pricing policy that requires limited information gathering can perform extremely well. Our results on policy complexity in Section 5 provide evidence that the policy of uniform pricing up to a cutoff state is superior according to a performance/complexity ratio measure.

The general state-dependent pricing strategy, where the price for a fill-in order arriving at time t is dependent on the state of the system at time t, is the most dynamic of the four policies we consider. Given the stochastic arrivals of core and fill-in customers, none of the customer groups can ex-ante predict what the state of the system will be. Although there may be some concern that such a dynamic pricing policy may turn away potential customers, we argue that existence of such fully state-dependent pricing policies may actually motivate some fill-in customers to become core customers. A parallel picture in the grocery industry is the emergence of loyalty cards where "loyal" (or long-term) customers are promised better deals than "walk-in" customers who have to pay the price they face during that week. In fact, it appears that supermarket chains are thus effectively competing with big discounters such as WalMart via using such customer loyalty programs and thus increasing their overall profitability (Patton 2002).

This research presents many opportunities for future work. For example, one could endogenize the price charged to core customers and optimally select both that price and fill-in prices. Furthermore, future research could endogenize the waiting time constraint as well. Additional future research could examine the impact of strategic competition on the pricing and waiting time policies for core and fill-in customers. Another opportunity exists in estimating actual demand functions for core and fill-in customers in a business-to-business setting and developing the corresponding optimal policies. Such an analysis would be more amenable to comparative statics as well. It is our hope that this research stimulates further interest in the integration of pricing and supply chain performance.

APPENDIX – PROOFS

Proof of Theorem 1.

(a) Part (a) follows directly from differentiating the Lagrangean for this problem.

(b) Sufficient conditions here follow from the Karush-Kuhn-Tucker sufficient conditions, i.e., $\lambda_f(p)p$ must be concave and $W_c - W_0$ must be convex. For linear demand, the former follows immediately. The later condition can be reduced to the condition $\frac{\partial^2 W_c}{\partial \lambda_f^2} \geq 0$, which can be easily verified. ♦

Proof of Lemma 1.

(a) From the balance equations for this Markov chain, it is straightforward to establish that $\pi_n = \left(\frac{\lambda_c}{\mu}\right)^{n-1}\left(\frac{\lambda_c + \lambda_f}{\mu}\right)\pi_0$. Solving for π_0 yields the result.

(b) Since core customer arrivals occur with a uniform rate in every system state, we can compute their expected wait as:

$$W_c = \sum_{i=0}^{\infty} \pi_i \frac{i+1}{\mu} .$$

The result follows from this and the results of part (a). ♦

Proof of Theorem 2.

(a) Part (a) follows directly from differentiating the Lagrangean for this problem.

(b) Here, we proceed by establishing that the Lagrangean (denoted by L) is quasi-concave. For this problem we have:

$$L = \pi_0 \lambda_f(p)p - \beta(W_c - W_0).$$

First, consider the case where:

$$\frac{1}{\mu - \lambda_c} = W_0 .$$

In this case, the only feasible choice of p is $p_{\max}$, where $\lambda_f(p_{\max}) = 0$. The remainder of this proof considers the case

$$\frac{1}{\mu-\lambda_c} < W_0 . \qquad \text{(A1)}$$

Substituting in the results of Lemma 1 and differentiating yields:

$$\frac{\partial L}{\partial p} = \frac{\mu-\lambda_c}{\left(\mu+\lambda_f(p)\right)^2}\left[-\lambda_f'(p)\left(\lambda_f(p)p+\frac{\beta}{\mu-\lambda_c}\right)+\left(\mu+\lambda_f(p)\right)\left(\lambda_f'(p)p+\lambda_f(p)\right)\right].$$

We proceed by investigating the behavior of this partial derivative. First, we look at extreme values of $p=0$ and $p=p_{\max}$ where $p_{\max}$ is defined by $\lambda_f(p_{\max})=0$

$$\left.\frac{\partial L}{\partial p}\right|_{p=0} = \frac{\mu-\lambda_c}{\left(\mu+\lambda_f(0)\right)^2}\left[-\lambda_f'(0)\left(\frac{\beta}{\mu-\lambda_c}\right)+\left(\mu+\lambda_f(0)\right)\left(\lambda_f(0)\right)\right] \geq 0,$$

and

$$\left.\frac{\partial L}{\partial p}\right|_{p=p_{\max}} = \frac{\mu-\lambda_c}{\mu^2}\left[-\lambda_f'(p_{\max})\left(\frac{\beta}{\mu-\lambda_c}\right)+\mu\lambda_f'(p_{\max})p_{\max}\right] \leq 0 .$$

The result at $p=p_{\max}$ follows from (A1): the waiting constraint cannot be binding at $p_{\max}$ and thus $\beta=0$.

Since $\frac{\mu-\lambda_c}{\left(\mu+\lambda_f(p)\right)^2} > 0$ holds by assumption on exogenous values, the roots and sign of $\frac{\partial L}{\partial p}$ are identical to the roots of $r(p)$, where:

$$r(p) = -\lambda_f'(p)\left(\lambda_f(p)p+\frac{\beta}{\mu-\lambda_c}\right)+\left(\mu+\lambda_f(p)\right)\left(\lambda_f'(p)p+\lambda_f(p)\right).$$

Now, under the assumption of linear demand, we have:

$$r'(p) = 2\left(\mu+\lambda_f(p)\right)\lambda_f'(p) < 0,$$

and thus $r(p)$ has a single root and L is quasi-concave. ♦

Proof of Lemma 2.

(a) From the balance equations for this Markov chain, it is straightforward to establish that

$$\pi_n = \begin{cases} \left(\frac{\lambda_c + \lambda_f}{\mu}\right)^n \pi_0 \text{ if } n \le s+1 \\ \frac{\left(\lambda_c + \lambda_f\right)^{s+1} \lambda_c^{n-(s+1)}}{\mu^n} \pi_0 \text{ if } n > s+1. \end{cases}$$

The result follows from solving for π_0 and summing π_0 through π_s.

(b) Since core customer arrivals occur with a uniform rate in every system state, we can compute their expected wait as:

$$W_c = \sum_{i=0}^{\infty} \pi_i \frac{i+1}{\mu} .$$

The result follows from this and the results of part (a). ♦

Proof of Theorem 3.

The Lagrangean (denoted by L) for this problem is:

$$L = F(s, p)\lambda_f(p)p - \beta(W_c(s, p) - W_0).$$

(a) If $\beta = 0$, we have:

$$L = F(s, p)\lambda_f(p)p,$$

and we can form the first-difference with respect to s:

$$\lambda_f(p)p(F(s+1, p) - F(s, p)).$$

This difference is not automatically positive because F is a distribution, because $F(s+1)$ and $F(s)$ are different distribution functions. However, it can be shown that:

$$F(s+1, p) - F(s, p) = \pi_0(s)\pi_0(s+1)\frac{\left(\frac{\lambda_c + \lambda_f(p)}{\mu}\right)}{\left(1 - \frac{\lambda_c}{\mu}\right)},$$

which is unambiguously positive, and therefore s^* is unbounded, the Lagrangean reduces to $L = \lambda_f(p)p$ and the result follows immediately. The first order conditions for the second case follow by differentiating the Lagrangean with respect to p and forming the first difference with respect to s.

(b) If $\beta \neq 0$, then the waiting time constraint is binding and we have $W_c = W_0$. To show that there exists an optimal, unique p for each value of s we need only to demonstrate that the equation $W_c = W_0$ yields a unique $p^*(s)$, which can be done by showing that W_c is strictly increasing in λ_f. While intuitively clear, this result can be demonstrated easily by showing that increasing λ_f results in a stationary distribution that exhibits first-order stochastic dominance and thus larger expected values of waiting time. ♦

Proof of Lemma 3.

(a) Follows using same logic as proof of Lemma 2. ♦

Proof of Theorem 4.

(a) If $\beta = 0$ the waiting time constraint can be ignored and we need only focus on maximizing the expected revenue earned per unit time. If the system is in state i, then the firm will earn nothing during the next state transition with probability $\frac{\lambda_c + \mu}{\lambda_{fi}(p_i) + \lambda_c + \mu}$ and will earn p_i with probability $\frac{\lambda_{fi}(p_i)}{\lambda_{fi}(p_i) + \lambda_c + \mu}$, for an expected earning of $\frac{\lambda_{fi}(p_i)p_i}{\lambda_{fi}(p_i) + \lambda_c + \mu}$. The rate at which state transitions are occurring is $\lambda_{fi}(p_i) + \lambda_c + \mu$, and thus we expect an earning rate of $\lambda_{fi}(p_i)p_i$. Since this same logic can be applied to every state, we conclude by symmetry that the optimal price for every state satisfies $\lambda'_{fj}(p_j^*)(p_j^*) + \lambda_{fj}(p_j^*) = 0$.

For $\beta \neq 0$, we begin with the Lagrangean (denoted by L) for this problem:

$$L = \sum_{i=0}^{\infty} \pi_i \lambda_{fi}(p_i) p_i - \beta(W_c - W_0).$$

Differentiating and collecting terms yields the set of first order conditions provided.

(b) For $\beta = 0$ the result follows immediately. ♦

Proof of Theorem 5.

(a) We denote the smallest price under which demand is zero as $p_{\max}$. If $\beta = 0$ all prices are equal from Theorem 4 and the result follows immediately. For $\beta \neq 0$, we note that the first-order conditions from part (a) are all of the form:

$$\lambda'_{fi}\left(p_i^*\right)\left[p_i^* - X_i\right] + \lambda_{fi}\left(p_i^*\right) = 0,$$

where X_i is given by:

$$\begin{aligned} X_i = &\left(\Pi - \lambda_{f,i+1}\left(p_{i+1}^*\right)p_{i+1}^* + \beta\left(\frac{i+2}{\mu} - W_c\right)\right)\frac{\partial\left(\pi_{i+1}/\pi_i\right)}{\partial\lambda_{fi}} + \\ &\left(\Pi - \lambda_{f,i+2}\left(p_{i+2}^*\right)p_{i+2}^* + \beta\left(\frac{i+3}{\mu} - W_c\right)\right)\frac{\partial\left(\pi_{i+2}/\pi_i\right)}{\partial\lambda_{fi}} + \\ &\left(\Pi - \lambda_{f,i+3}\left(p_{i+3}^*\right)p_{i+3}^* + \beta\left(\frac{i+4}{\mu} - W_c\right)\right)\frac{\partial\left(\pi_{i+3}/\pi_i\right)}{\partial\lambda_{fi}} + \ \ldots \end{aligned}$$

By inspection of the form of X_i, it is clear that for any $\beta > 0$, $\lim_{i\to\infty} X_i = \infty$ for any allowed values of the decision variables. It is also straightforward to show that $\frac{\partial p_i^*}{\partial X_i} = \frac{1}{2}$ under linear demand, and thus p_i is increasing in X_i. We employ backwards recursion to prove the monotonicity result. First, note that we can form the following recursion for the X_i's:

$$X_i = \left(\Pi - \lambda_{f,i+1}\left(p_{i+1}^*\right)p_{i+1}^* + \beta\left(\frac{i+2}{\mu} - W_c\right)\right)\frac{1}{\mu} + \left(\frac{\lambda_{f,i+1} + \lambda_c}{\mu}\right)X_{i+1}.$$

Begin by assuming $X_i - X_{i-1} \geq 0$. Now, examine $X_{i-1} - X_{i-2}$:

$$X_{i-1} - X_{i-2} = X_{i-1}\left(1 - \frac{\lambda_c}{\mu}\right) + \frac{1}{\mu}\left(\lambda_{f,i-1}\left(p_{i-1}^*\right)\left(p_{i-1}^* - X_{i-1}\right) - \left(\Pi + \beta\left(\frac{i}{\mu} - W_c\right)\right)\right).$$

Now, we can rewrite $X_i - X_{i-1} \geq 0$ as:

$$X_{i-1} \geq \left(\frac{\mu}{\mu - \lambda_c}\right)\left(\Pi - \lambda_{f,i}\left(p_i^*\right)\left(p_i^* - X_i\right) + \beta\left(\frac{i+1}{\mu} - W_c\right)\right)\frac{1}{\mu},$$

so we have:

$$X_{i-1} - X_{i-2} \geq \frac{1}{\mu}\left(\frac{\beta}{\mu} - \lambda_{f,i}\left(p_i^*\right)\left(p_i^* - X_i\right) + \lambda_{f,i-1}\left(p_i^*\right)\left(p_i^* - X_{i-1}\right)\right).$$

Since, $X_i - X_{i-1} \geq 0$ by assumption, we have $\lambda_{f,i}\left(p_i^*\right)\left(p_i^* - X_i\right) \leq \lambda_{f,i-1}\left(p_{i-1}^*\right)\left(p_{i-1}^* - X_{i-1}\right)$ by the envelope theorem, and thus $X_{i-1} - X_{i-2} \geq 0$. We have not shown a base case for this recursion, but from these results we can conclude that the X_i's are unimodal. Since we have $\lim_{i\to\infty} X_i = \infty$, either all X_i's are infinite or the X_i's are increasing. It follows that $p_i^* \leq p_{i+1}^*$ for all i. ♦

REFERENCES

Adiri, I., & Yechiali, U. (1974). Optimal Priority-Purchasing and Price Decisions in Nonmonopoly and Monopoly Queues. *Operations Research, 22*, 1051-1066.

Arvind Rajan, R., & Steinberg, R. (1992). Dynamic Pricing and Ordering by a Monopolist. *Management Science, 38*, 240-262.

At What Price? Guidelines for a Customer-Focused Pricing Strategy. (2000). Retail Industry Report: Arthur Andersen.

Baker, W., Marn, M., & Zawada, C. (2001). Price Smarter on the Net. *Harvard Business Review, 79*(2), 122-127.

Balachandran, K. R. (1972). Purchasing Priorities in Queues. *Management Science, 18*(5), 319-326.

Bertsekas, D. P. (1987). *Dynamic Programming: Deterministic and Stochastic Models*. Englewood Cliffs, NJ: Prentice Hall.

Biller, S., Chan, L. M. A., Simchi-Levi, D., & Swann, J. (2002). *Dynamic Pricing and the Direct-to-Customer Model in the Automotive Industry.* Unpublished Working Paper, Georgia Institute of Technology, Atlanta.

Chan, L. M. A., Simchi-Levi, D., & Swann, J. (2001). *Dynamic Pricing Strategies for Manufacturing with Stochastic Demand and Discretionary Sales.* Unpublished Working Paper, Georgia Institute of Technology, Atlanta.

Chen, X., & Simchi-Levi, D. (2002). *Coordinating inventory control and pricing strategies with random demand and fixed ordering cost: the finite horizon case.* Unpublished Working Paper.

Cisco Thought Leadership Summit (2001), "Real Time Profit Optimization: Coordinating Demand and Supply Chain," *Tuck-Cisco White Paper*, The Tuck School of Business, Hanover, NH.

Cover, T. M. and J. A. Thomas (1991), Elements of Information Theory, Wiley-Interscience.

Dewan, S., & Mendelson, H. (1990). User Delay Costs and Internal Pricing for a Service Facility. *Management Science, 36*(12), 1502-1517.

Eliashberg, J., & Steinberg, R. (1987). Marketing-Production Decisions in an Industrial Channel of Distribution. *Management Science, 33*(8), 981-1000.

Eliashberg, J., & Steinberg, R. (1991). Marketing-Production Joint Decision Making. In J. E. a. J.D.Lilien (Ed.), *Management Science in Marketing*. Amsterdam: North Holland.

Federgruen, A., & Heching, A. (1999). Combined Pricing and Inventory Control Under Uncertainty. *Operations Research, 47*(3), 454-475.

Fleischmann, M., Hall, J. M., & Pyke, D. F. (2004). Smart Pricing: A review of recent, and some seminal, work linking pricing decisions with operational insights. *Sloan Management Review, Forthcoming*(Winter).

Gijsbrechts, E. (1993). Prices and Pricing Research in Consumer Marketing: Some Recent Developments. *International Journal of Research in Marketing, 10*(2), 115-151.

Gilbert, S. M. (2000). Coordination of Pricing and Multiple-Period Production Across Multiple Constant Priced Goods. *Management Science, 46*(12), 1602-1616.

Gross, D., & Harris. (1985). *Fundamentals of Queueing Theory* (Second Edition ed.). New York: John Wiley & Sons.

Hall, J. M., Kopalle, P. K., & Krishna, A. (2002). *A Multi-Product Model of Retailer's Dynamic Pricing and Ordering Decisions: Normative and Empirical Analysis.* Unpublished Working Paper, Dartmouth College, Hanover, NH.

Karlin, S., & Carr, C. R. (1962). Prices and Optimal Inventory Policy. In K. Arrow & S. Karlin & H. Scarf (Eds.), *Studies in Applied Probability and Management Science*. Stanford, California: Stanford University Press.

Kim, Y. J., & Mannino, M. V. (2003). Optimal incentive-compatible pricing for M/G/1 queues. *Operations Research Letters, 31*(6), 459-461.

Kopalle, P., Mela, & Marsh. (1999). The Dynamic Effect of Discounting on Baseline Sales: Empirical Analysis and Normative Pricing Implications. *Marketing Science, 18*(3), 317-332.

Kopalle, P. K., Rao, A. G., & Assunção, J. L. (1996). Asymmetric Reference Price Effects and Dynamic Pricing Policies. *Marketing Science, 15*(1), 60-85.

Kunreuther, H., & Richard, J. F. (1971). Optimal Pricing and Inventory Decisions for Non-Seasonal Items. *Econometrica, 39*, 173-175.

Kunreuther, H., & Schrage, L. (1973). Joint Pricing and Inventory Decisions for Stable Priced Items. *Management Science, 19*(7), 732-738.

Lau, A. H.-L., & Lau, H.-S. (1988). The Newsboy Problem with Price-Dependent Demand Distribution. *IIE Transactions, 20*(2), 168-175.

Lester, T. (2002, January 30). Inside Track: How to ensure that the price is exactly right. *Financial Times*.

Mendelson, H. (1985). Pricing Computer Services: Queueing Effects. *Communications of the ACM, 28*(3), 312-321.

Mendelson, H., & Whang, S. (1990). Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue. *Operations Research, 38*(5), 870-883.

Naor, P. (1969). The Regulation of Queue Size by Levying Tolls. *Econometrica, 37*(1), 15-24.

Noble, P. M., & Gruca, T. S. (1999). Industrial Pricing: Theory and Managerial Practice. *Marketing Science, 18*(3), 435-454.

Patton, S. (2002), "Food Fight," *CIO*, 16 (2), 86-92.

Pekelman, D. (1974). Simultaneous Price-Production Decisions. *Operations Research, 22*(4), 788-794.

Plambeck, E. L. (2000). *Pricing, Leadtime Quotation and Scheduling in a Queue with Heterogeneous Customers.* Unpublished Working Paper, Stanford University, Stanford, CA.

Raman, K., & Chatterjee, R. (1995). Optimal Monopolist Pricing Under Demand Uncertainty in Dynamic Markets. *Management Science, 41*(1), 144-162.

Reda S. (2002), Retailers Slow to Adopt Analytics Software. *Stores*, 84(6), 22.

Retail Revenue Management. (2001). The Forrester Report, Forrester Research, Inc.

Rump, C. M., & Stidham Jr., S. (1998). Stability and Chaos in Input Pricing for a Service Facility with Adaptive Customer Response to Congestion. *Management Science, 44*(2), 246-261.

Seifert, R. W., Thonemann, U. W., & Hausman, W. H. (2004). Optimal procurement strategies for online spot markets. *European Journal of Operational Research, 152*(3), 781-799.

Sogomonian, A. G., & Tang, C. S. (1993). A Modeling Framework for Coordinating Promotion and Production Decisions within a Firm. *Management Science, 39*(2), 191-203.

Stidham Jr., S. (1985). Optimal Control of Admission to a Queueing System. *IEEE Transactions on Automatic Control, AC-30*(8), 705-713.

Stidham Jr., S. (1992). Pricing and Capacity Decisions for a Service Facility: Stability and Multiple Local Optima. *Management Science, 38*(8), 1121-1139.

Thomas, J. (1970). Price-Production Decisions with Deterministic Demand. *Management Science, 16*(11), 747-750.

Thomas, L. J. (1974). Price and Production Decisions with Random Demand. *Operations Research, 22*(3), 513-518.

Wagner, H., & Whitin, T. M. (1958a). Dynamic Problems in the Theory of the Firm. *Naval Research Logistics, 5*, 53-74.

Wagner, H., & Whitin, T. M. (1958b). Dynamic Version of the Economic Lot Size Model. *Management Science, 5*(1), 89-96.

Whitin, T. M. (1955). Inventory Control and Price Theory. *Management Science, 2*(1), 61-68.

State i	p_i^*	$\lambda_{fi}(p_i^*)$
0	760.73	23.93
1	856.12	14.39
2	902.82	9.72
3	930.55	6.94
4	949.22	5.08
5	962.99	3.70
6	973.94	2.61
7	983.11	1.69
8	991.39	0.86
9	999.27	0.07
10 and higher	1000	0

Table 1. Results for General State-Dependent Pricing Example.

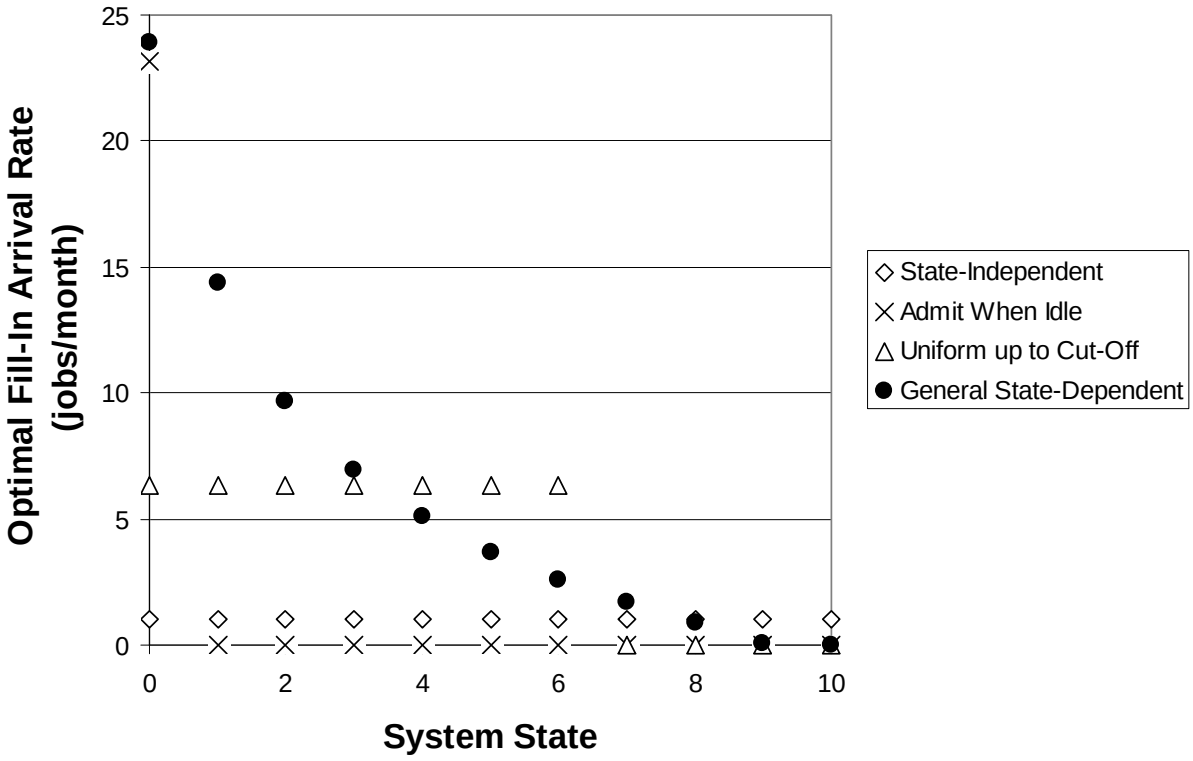

Figure 1. Mean Fill-in Arrival Rate as a Function of the System State.

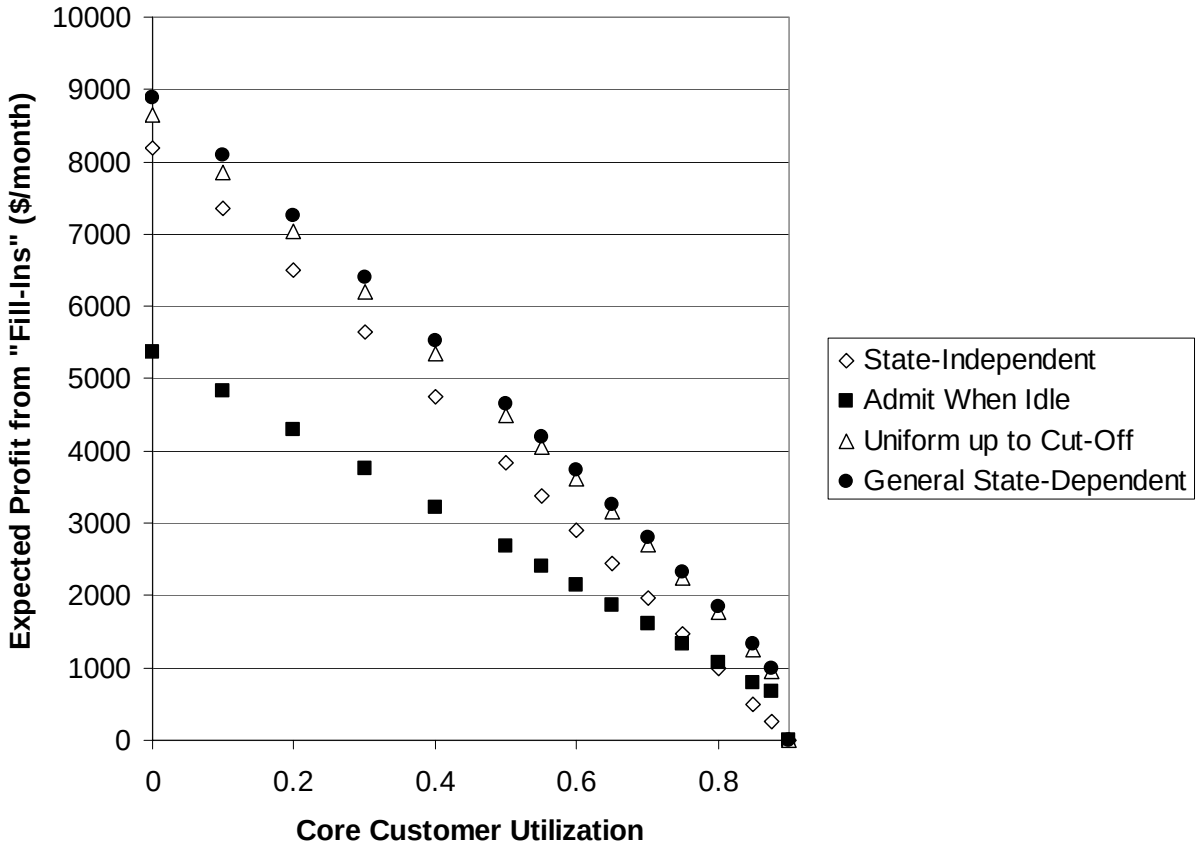

Figure 2. Numerical Examples for Large Market Demand Function

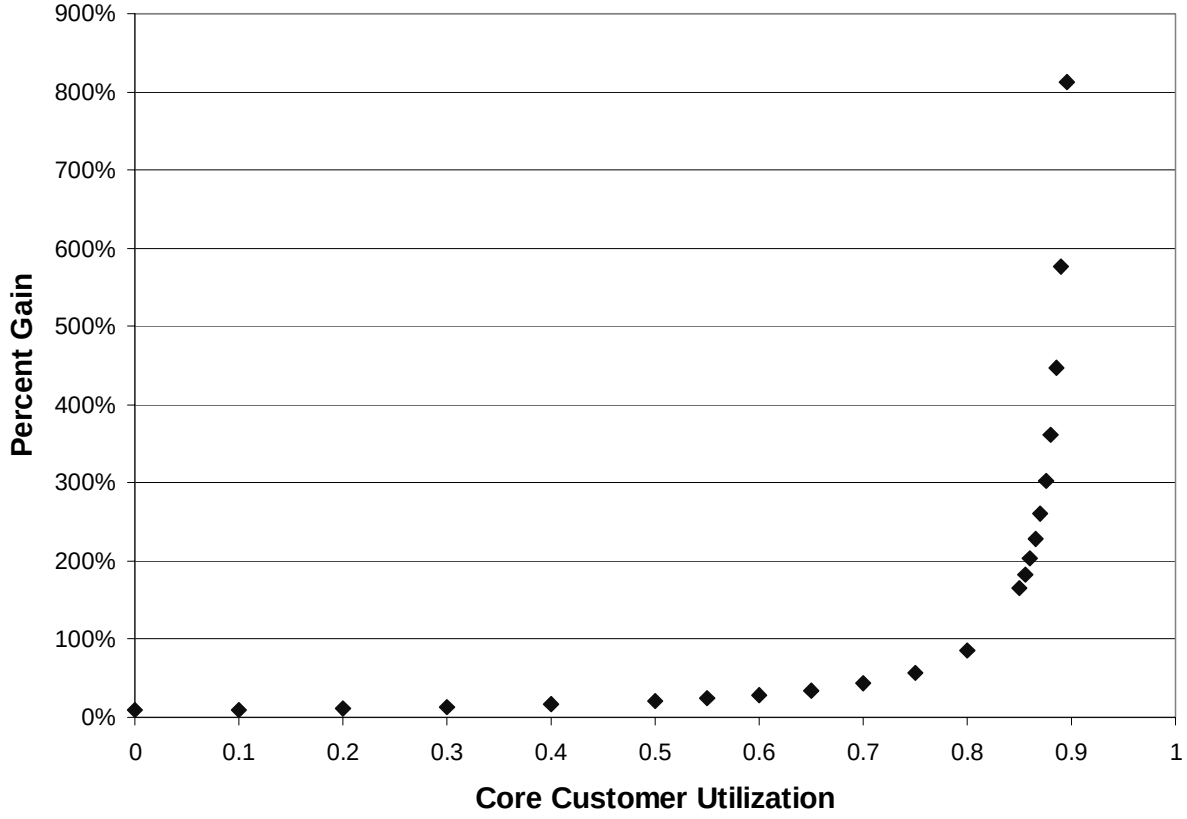

Figure 3. Relative Gain From Using Internal State Information Relative to State Independent Pricing (Large Market).

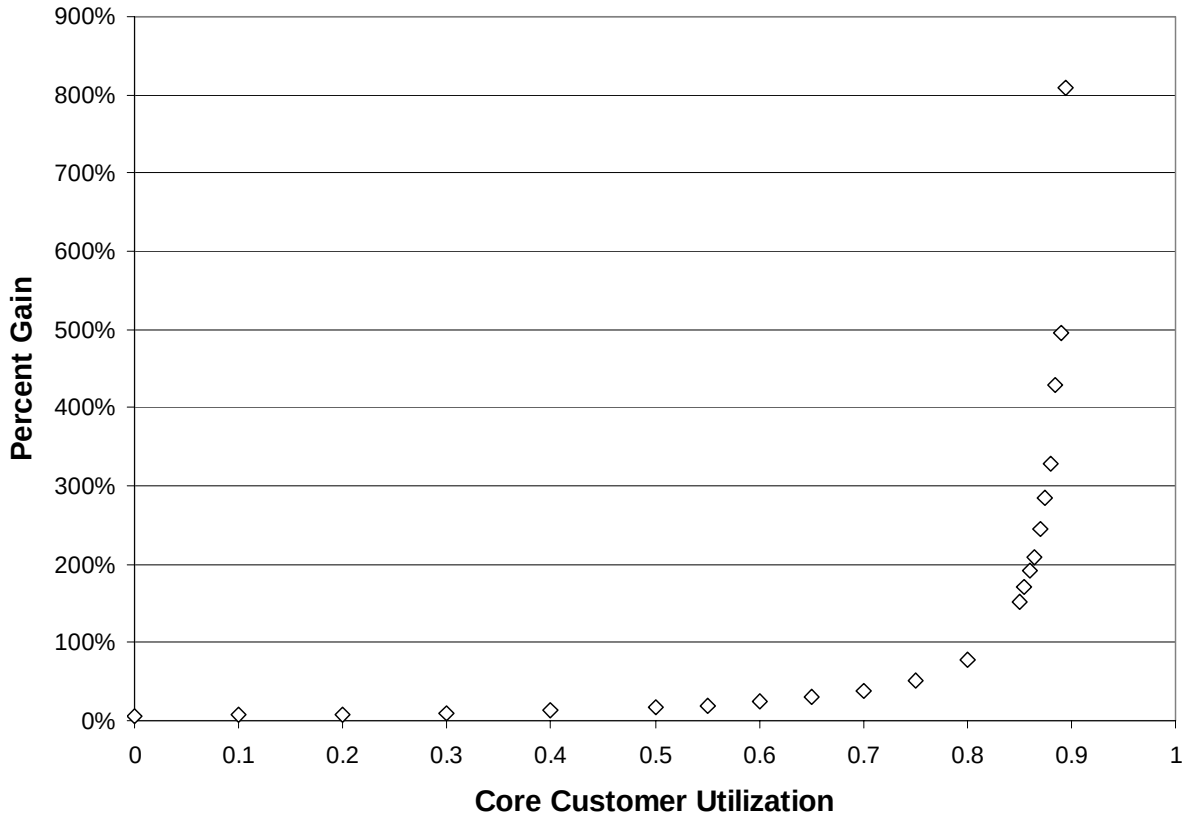

Figure 4. Relative Gain From Uniform Price up to Cut-off Policy vs. State-Independent Pricing (Large Market).

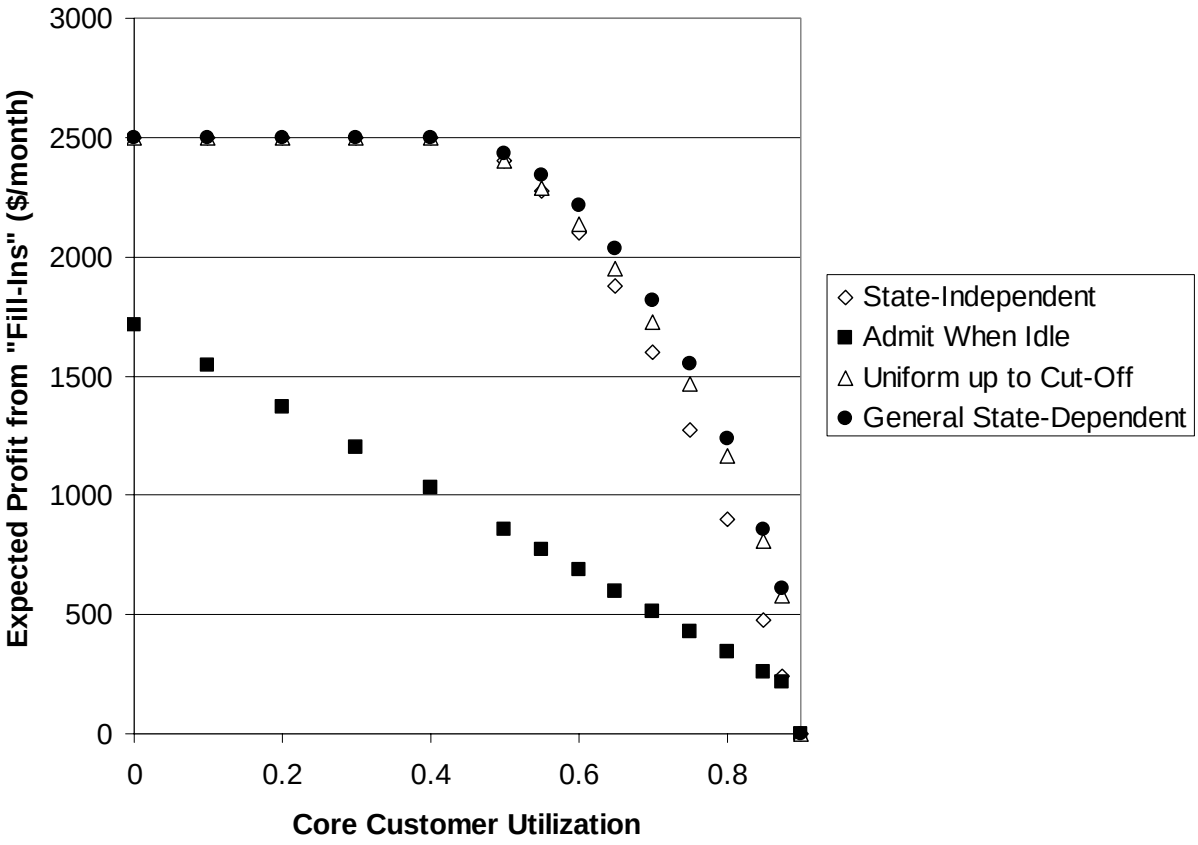

Figure 5. Numerical Examples for Small Market Demand Function

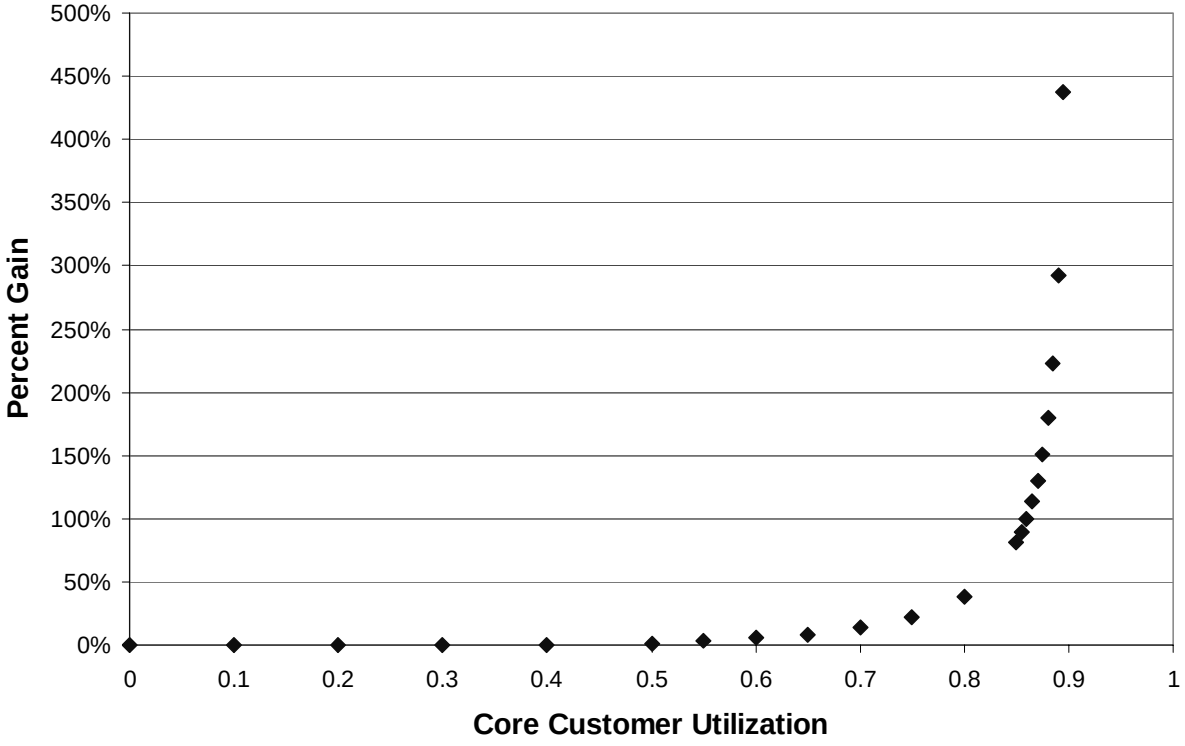

Figure 6. Relative Gain From Using Internal State Information (Small Market).

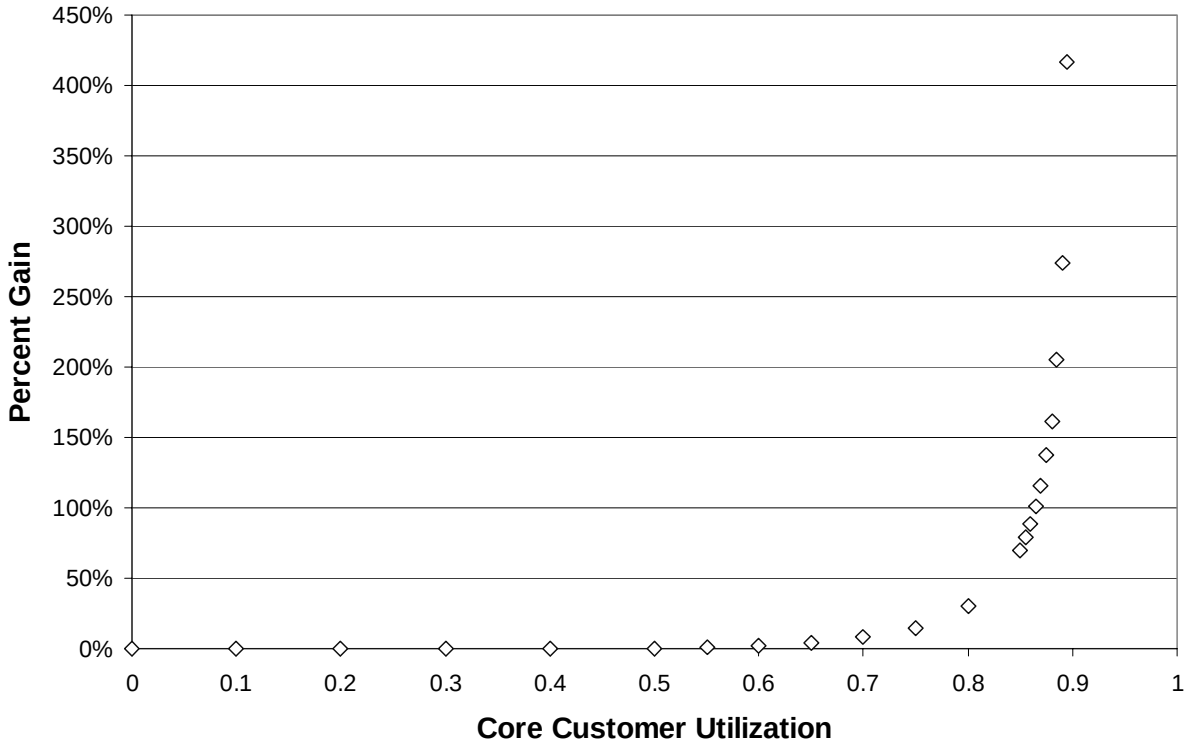

Figure 7. Relative Gain From Uniform Price up to Cut-off Policy vs. State-Independent Pricing (Small Market)

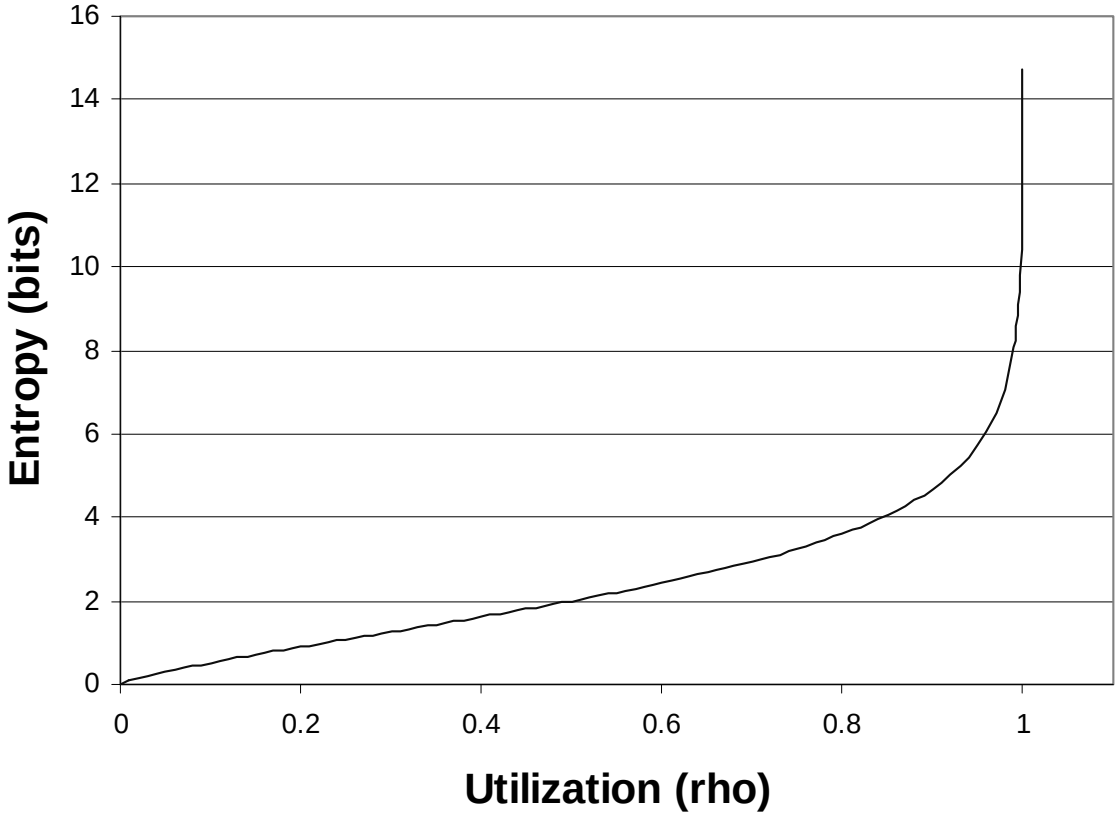

Figure 8. M/M/1 State Entropy vs. Utilization.

Policy:	State-Independent	Admit When Idle	Uniform up to Cutoff	General State-Dependent
Expected Gain From "Fill-Ins" ($/month)	990	1073	1767	1840
Entropy of Pricing Signal (bits)	0	0.329	0.880	4.172
Additional Return per Bit ($/bit)	-	252	883	204

Table 2. Calculation of "Return per Bit."