Assessing the Benefits of Different Stock-Allocation Policies for a Make-to-Stock Production System

Francis de Véricourt • Fikri Karaesmen • Yves Dallery
Fuqua School of Business, Duke University, Durham, North Carolina 27708
Laboratoire Productique Logistique, École Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry Cedex, France
Laboratoire Productique Logistique, École Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry Cedex, France
fdv1@mail.duke.edu • fikri@pl.ecp.fr • dallery@pl.ecp.fr

We consider a manufacturing facility that produces a single item that is demanded by several different classes of customers. The inventory-related cost performance of such a system can be improved by effective allocation of production and inventories. We obtain the optimal parameters for three easily implementable allocation policies. Our results cover the case of linear backorder costs as well as fill-rate constraints. We compare the optimal performance of these control policies to gain insights into the benefits of different production and stock-allocation rules.
(*Inventory/Production: Optimal Policies, Stock Allocation; Queues: Make-to-Stock System*)

1. Introduction

In this article, we investigate stock-rationing problems in a manufacturing environment. Because limited production capacity is an important characteristic, stock-allocation problems naturally arise in this setting. Our objective is to investigate the effects of different stock-rationing policies on inventory-related costs and service levels, within a framework that addresses uncertainties in demand and in manufacturing as well as the effects of limited capacity. One appropriate framework is that of queuing-based inventory models that have led to a unified treatment of several central issues in production-inventory theory (Buzacott and Shanthikumar 1993). Our investigation of the effects of stock rationing is therefore based on a model within that framework: the make-to-stock queue with several classes of clients.

In our model, a production facility produces a single item in a make-to-stock mode. The same item is demanded by several classes of customers that may differ in their demand rates, backorder costs, or service levels. When a demand arrives, depending on its class, it may be satisfied immediately from stock (when available) or may be back-ordered to be satisfied later. Because of the differences in backorder costs to customers, it is possible that a demand is made to wait, in consideration of future, more expensive arrivals, even though there may be items available in stock.

To investigate the effects of different production and stock-allocation policies in a capacitated production setting, our analysis focuses on three different policies. The first policy is a base-stock policy using a simple FCFS rule for stock allocation. It is simple to communicate, is frequently used in practice, and is optimal when different demand classes have identical backorder costs. The other two policies use priority allocation and privilege customers with higher waiting costs in allocation decisions. The difference between these two policies is the additional feature of

1523-4614/01/0302/0105$05.00
1526-5498 electronic ISSN

the reservation of inventories for future demand arrivals (with high backorder costs). The third policy is a generalization of the second policy, which does not reserve inventories. In fact, it can be shown that this last policy is optimal under certain technical assumptions; see de Véricourt et al. (2000b). This optimality comes, however, at the expense of additional parameters to optimize.

Our initial contribution is to compute the optimal levels of the parameters that minimize inventory and backorder costs for two of the heuristic policies considered. It turns out that the optimal levels are easily expressed in terms of the system parameters, leading immediately to simple insights on the relative benefits of each policy. Combining this with the results in de Véricourt et al. (2000b) pertaining to the parameters of the optimal policy, enables us to compare, through numerical experiments, the optimal performance within each class.

To complement our numerical results, we analyze the three policies in detail in the case of high utilization, which is the relevant regime in certain manufacturing environments such as the semiconductor industry. Lastly, we discuss the issues of parameter optimization and its implications on performance when backorder costs are replaced by fill-rate constraints. Overall, the analysis enables us to present a rather complete picture of the managerial implications of stock-rationing issues.

The paper is structured as follows: A review of the relevant literature is presented in §2. Section 3 gives the mathematical formulation of the problem and introduces the particular control policies that we investigate. Section 4 provides a performance comparison. In §5, we analyze an important special case: a heavily loaded system. Section 6 extends the formulation and the results to the case of service-level constraints. Our conclusions are presented in §7.

2. Literature Review

Besides the earlier-mentioned applications in product-variety management that have gained significance as a result of increasing product proliferation, stock allocation is a question that naturally arises in several models of classical inventory theory. Jackson (1988) and Mc Gavin et al. (1993) study optimal stock-allocation problems for periodic review systems. Although the literature in this domain is relatively rich because of the nature of the assumptions of uncapacitated production, constant lead times and identical clients, most of the research in this area does not directly address the issues of stock rationing.

The inventory models that directly investigate the issues of stock rationing fall into the *single-location multiple-customer* category. In pioneering work, Topkis (1968) has characterized the structure of optimal ordering and rationing policies for such a model. In this model, the optimal ordering policy is a base-stock policy and the optimal amounts of stocks reserved are determined by time-dependent threshold values.

In other related work based on stochastic inventory models, Nahmias and Demmy (1981) study several inventory models where rationing is relevant and compare the inventory-related cost and service-level performance of systems that employ rationing with those that do not. Their results indicate that rationing stocks improves performance significantly. Cohen et al. (1988) study the performance of a system operating under a (s, S) policy. There are two classes of customers and one class has strict priority over the other. Finally, Frank et al. (1999) study a problem with two classes of customers, in which the demands of the first class have to be satisfied, but the second-class demands can be rejected. They partially characterize the optimal policy and propose heuristic control policies that have close-to-optimal performance. All of the above articles treat interesting aspects of stock rationing, but they do not explicitly capture the effects of limited production capacity, which is central to our formulation.

More recently, a number of production-inventory problems for capacitated systems involving multiple customer classes have been studied in the context of the make-to-stock queue. The rationing problem in this setting resembles a closely related category of multi-item scheduling problems. In this latter category of problems, multiple classes of customers demand different products from the manufacturing system that has to schedule production by dynamically

sharing capacity. Wein (1992), Veatch and Wein (1996), and Pena-Perez and Zipkin (1997) proposed heuristic solutions to this problem. In subsequent work, Ha (1997a) partially characterized the structure of the optimal policy for two classes of customers, and de Véricourt et al. (2000a) obtained a sharper characterization of this structure. The model studied in this paper can be viewed as a *standardized* version of the above models where only a single product type is stored.

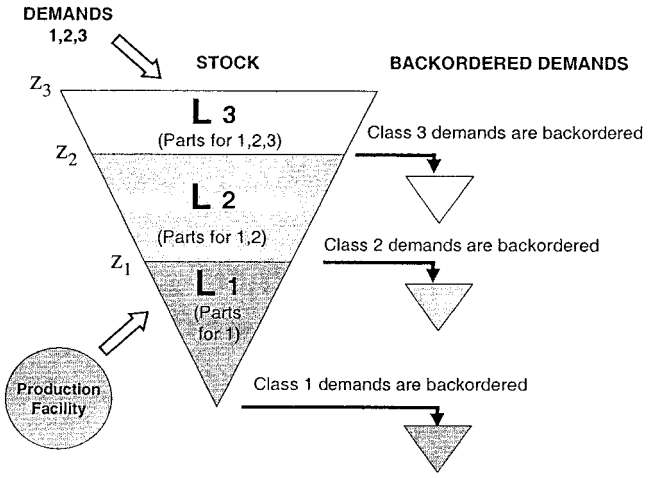
The rationing problem for the make-to-stock queue with multiple demand classes was first studied by Ha (1997b and 1997c). Ha (1997b) formulates and studies the optimal rationing and production control of a multiclass system with lost sales. He shows that the optimal production-control policy is a base-stock policy, and the optimal rationing policy is described by threshold levels corresponding to the different demand classes. When the stock on hand is above the threshold level of a certain class of demand, it is satisfied from on hand stock, and otherwise it is lost. In addition, this multiple-threshold rationing policy performs significantly better than policies that do not employ rationing.

When backorders are allowed, for a complete state description it is necessary to keep track of the levels of the backorder queues of each customer class. The above problem then becomes significantly more difficult because the dimension of the corresponding optimal control problem increases. This case is studied in Ha (1997c), who shows that the optimal production-control policy is still a base-stock policy and that the optimal rationing policy has a monotone structure. In parallel work, we establish a complete characterization of the optimal stock-rationing policy (de Véricourt et al. 2000b) for this case. The optimal policy turns out to be surprisingly intuitive: As in the lost-sales case (Ha 1997a), there are thresholds for each product such that it is optimal to satisfy the arriving demand from a customer from the on-hand stock if the stock level is above the threshold for that customer. Moreover, these thresholds determine production priorities for backordered products in a simple way.

Even though the previous results have shed light onto the structure of optimal rationing and production-control policies, the potential benefits of stock rationing cannot be entirely understood without a complete investigation. Earlier works of Nahmias and Demmy (1981) and Ha (1997b) indicate that stock rationing may have significant benefits in terms of inventory-related costs, respectively, for uncapacitated inventory models and for production-inventory systems with lost sales. However, the comparisons made in the preceding models have their drawbacks: Nahmias and Demmy do not compare optimal rationing with optimal nonrationing policies because of the difficulty of optimization. Ha's comparisons, because of the lost-sales assumption, overlook the effect of critical regimes where production capacity is barely enough to meet the average demand. We complement and extend the results of both articles here by treating multiple classes of customers in a rather general setting. In addition, most of our results are expressed in simple formulas from which several simple insights into the relative benefits of optimal production and stock allocation can be drawn.

To summarize and position the contributions of this paper with respect to closely related existing work, note that there are several equally important issues in the analysis of dynamic optimization problems of production-inventory systems. An important direction of analysis is the investigation of the structure of optimal control policies. Ha (1997c) and de Véricourt et al. (2000a) are examples that analyze the stock rationing (i.e., single-item/multiclient) and production scheduling (i.e., multi-item) problems for the make-to-stock queue, respectively. This type of investigation usually gives partial answers about what the optimal control policy should be but does not provide a complete explicit solution. Only in certain special cases is the structural characterization simple and precise enough to lead to tractable optimal policies. This is typically the case where the state space of the inventory-backlog process is one-dimensional as in the single-class make-to-stock queue or in the inventory-rationing problem with lost sales (Ha 1997b). When the state space of the inventory-backlog processes has to be represented by several variables, the only explicit characterization seems to be that in de

Figure 1 The ML Policy for $n = 3$



Véricourt et al. (2000b) for the rationing problem with backorders considered here.

Because the structure of optimal policies in itself does not provide insights into how optimality reflects into cost savings with respect to other (suboptimal) policies, a parallel and complementary direction of analysis investigates the issues of relative performances of heuristic policies. Veatch and Wein (1996), Pena-Perez and Zipkin (1997), and Ha (1997b) comprise contributions in this sense. This paper follows a similar direction. Given that the optimal inventory-rationing policy has been characterized (de Véricourt et al. 2000b), the objective here is to gain insights into the relative benefits of using the optimal policy over other policies, which are attractive because of their simplicity. To achieve this end, we present the optimal parameters of the other two policies, which are of interest themselves. We then carry out a comparative analytical and numerical investigation that identifies the conditions under which using the optimal policy is worthwhile in terms of cost savings. Finally, all related works cited above favor a backorder or lost-sales cost formulation. We provide explicit analytical results for a formulation where backorder costs are replaced by service-level constraints. This second formulation seems to be the more relevant one in many environments.

3. Stock-Allocation and the Control Policies

3.1. Formulation of the Model

Consider a production facility that produces a single product to stock. Each finished item is placed in the finished-goods inventory. There are n classes of customers for this product. When the on-hand inventory is zero, demands are back-ordered. When it is positive, an arriving demand can be either satisfied by the on-hand inventory or can be backordered. We consider a holding cost h (per unit, per time) and backorder costs b_i (per unit, per time) for class i customers. Without loss of generality we assume that the backorder costs are ordered such that $b_1 > \cdots > b_n$ (if $b_i = b_j$, the classes i and j can be considered as a single class with arrival rate $\lambda_i + \lambda_j$ and backlog cost $b_i = b_j$). The customers of class i arrive according to a Poisson process with rate λ_i. Let $\lambda = \Sigma_{i=1}^{n} \lambda_i$. The production time is exponentially distributed with mean $1/\mu$. We also define $\rho = \lambda/\mu$, the traffic intensity of the system and $\rho_k = (\Sigma_{i=1}^{k} \lambda_i)/\mu$, the traffic intensity of the subsystem comprising the first k classes (we have accordingly $\rho_n = \rho$, and we take $\rho_0 = 0$). To ensure the stability of the system, we assume that $\rho < 1$.

The state of the system can be described by the vector $\mathbf{x}(t) = (x_0(t), x_1(t), \ldots, x_n(t))$, with $x_i \in N$, $0 \le i \le n$. $x_0(t)$ is the on-hand inventory at time t and $x_i(t)$ with $1 \le i \le n$ is the number of backorders at time t for class i. $\mathbf{X}(t) = (X_0(t), X_1(t), \ldots, X_n(t))$ denotes the associated random variables.

The system operates in a make-to-stock-type environment where inventories are built in advance in anticipation of future demands. In addition, because customer classes differ in their backorder costs, a stock-allocation problem arises. Assume that a class i demand arrives; should it be immediately satisfied from the inventory, or should it be back-ordered so that the inventory is saved for future demands of classes $1, \ldots, i-1$? Both production and allocation decisions obviously depend on the state of the system: the current on-hand inventory and the backorders of each class. An appropriate control policy must then specify the decisions of:

(1) Production—Whether to produce an item or not;

(2) Allocation—

- Production—Whenever the production of an item is completed, whether to use this item to reduce the number of backorders of a class, or to increase on-hand inventory,
- Stocks—When a demand occurs, whether to satisfy it from the on-hand inventory, or to backorder it.

(A formal definition of a control policy is presented in Appendix 1).

When in state $\mathbf{x}$, the system incurs a cost rate $c(\mathbf{x})$ that is equal to

$$c(\mathbf{x}) = hx_0 + \sum_{i=1}^{n} b_i x_i.$$

Our objective is to find a control policy π that minimizes the expected average cost over an infinite horizon:

$$\min_{\pi} \lim_{T\to\infty} \frac{1}{T} E^{\pi}\left[\int_0^T c(\mathbf{X}(t))\, dt\right]. \tag{1}$$

The optimal allocation of production and inventories is then modeled as a multidimensional stochastic-control problem. In de Véricourt et al. (2000b), it is shown that under certain conditions the policy that minimizes (1) can be characterized. This optimal policy has a multilevel structure; hence, we refer to it as the ML policy. The presentation of the ML policy is deferred to the next section.

Although the ML policy is optimal, there are several other plausible allocation policies. To gain an understanding of the value of optimal inventory allocation, we consider two alternatives. These policies not only provide a benchmark for the ML policy, but also have the merit of possessing fewer parameters to optimize. Furthermore, they have close-to-optimal performance under certain conditions. The common point of all three policies is that in terms of the production decisions they are members of the base-stock family, which drives the system to a target base-stock level. Because the underlying difference between the policies is the way the system is driven towards its base-stock level, we differentiate them by referring to the respective production and stock-allocation policies and omit the *base-stock* term in the description of the policy for simplicity. As a result of this shortcut, the *First-Come-First-Served* (*FCFS*) *Policy,* for instance, refers to a base-stock policy with FCFS allocation of production and stocks.

3.2. The First-Come-First-Served (FCFS) Policy

The FCFS policy takes the allocation decisions with respect to the order of arrival of demands. It is described by a single parameter, a base-stock level z (≥ 0). At this point, no claims are made for the optimality of a base-stock policy if customers are satisfied FCFS. Nevertheless, a base-stock policy is simple and reasonable. Typically, the system starts with an on-hand inventory level equal to z. The controls of a FCFS policy are then:

(1) Production—Produce if and only if $x_0 < z$ or backlogs exist;

(2) Allocation—

- Production—If there are backordered demands, satisfy them in the order of their arrival (regardless of their class). If there are no backorders, add produced items to the on-hand inventory.
- Stocks—Satisfy arriving demand regardless of its class if the on-hand inventory is not empty, $x_0 > 0$; back-order it otherwise.

There are several reasons for considering this policy. First, in our experience, it seems to be common industrial practice. Second, it is the prevailing assumption in the multiretailer-inventory literature. Finally, it provides a benchmark for the performance of any policy that does not differentiate customers by their class. It is also interesting to note that if customers were identical in their backorder costs, this allocation policy would be optimal (just as any other ''nonidling when backordered'' policy).

The optimal base-stock level $\hat{z}$ and the optimal average cost, g_{fcfs}, of the FCFS policy or an n-class problem are given by the following property:

PROPERTY 1. *The optimal FCFS policy of an n-class problem is characterized by the base-stock level equal to*

$$\hat{z} = \left\lfloor \frac{\ln \frac{h}{\hat{b} + h}}{\ln \rho} \right\rfloor,$$

where $\hat{b}$ *is the aggregate backorder cost*

$$\hat{b} = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} b_i.$$

The optimal cost is then given by

$$g_{fcfs} = \hat{b} \frac{\rho^{\hat{z}+1}}{1 - \rho} + h\left[\hat{z} - \frac{\rho}{1 - \rho}(1 - \rho^{\hat{z}})\right].$$

The proof can be found in Appendix 2.

REMARK. The proof of Property 1 uses the fact that the backorder queues of each class can be viewed as a single backorder queue with aggregate backorder cost $\hat{b}$. For this equivalent single-class system, a base-stock policy (with $z \geq 0$) is optimal, thereby suggesting the optimality of a base-stock policy for the multiclass FCFS system.

3.3. The Strict Priority Policy

When there are backorders, a FCFS policy satisfies the demands in the order of their arrival, which, in general, is not optimal. One way to improve the performance of the system is to allocate production more efficiently when satisfying backordered demands. For a corresponding pure make-to-order system, a $c\mu$ rule is optimal (Baras et al. 1985), which in our context implies producing to satisfy the backorders of the class with the largest b in priority. The policy we will present exploits this property.

A Strict Priority (SP) policy is characterized by a base-stock level z. To facilitate the presentation, let us also define the function $m(x)$ that represents the class with the highest unit backorder cost among all backlogged classes. Because $b_1 > \cdots > b_n$, $m(\mathbf{x}) = \min_{i:x_i>0}(i)$. The controls of an SP policy are then:

(1) Production—Produce if and only if $x_0 < z$ or backlogs exist.

(2) Allocation—

- Production—If there are backlogs, allocate the item to class $m(\mathbf{x})$. Otherwise, put the item into on-hand inventory.
- Stocks—Satisfy arriving demand regardless of its class if the on-hand inventory is not empty ($x_0 > 0$); backorder it otherwise.

The recurrent states are the same as those of the FCFS policy. In fact, as long as the on-hand inventory is not zero (as long as $x_0 > 0$), the SP policy makes the same decisions as the FCFS policy. On the other hand, in stockout situations customers of class i are given allocation priority over customers of classes $i + 1, i + 2, \ldots, n$.

As in the FCFS policy, the optimal SP policy seeks a trade-off between two cost parameters h, the unit inventory cost, and $\tilde{c}$, an equivalent (aggregate) backorder cost. Unlike in the FCFS policy, the equivalent unit backorder cost depends on the particular production-capacity allocation. The key to the property below, however, is that this cost is independent of the choice of the base-stock level. The optimal base-stock level $\tilde{z}$, and the optimal average cost g_{sp} of the SP Policy for an n-class problem are given by the following property where $\tilde{c}$ is the aggregate backorder cost mentioned above.

PROPERTY 2. *The optimal SP policy of an n-class problem is characterized by a base-stock level equal to:*

$$\tilde{z} = \left\lfloor \frac{\ln \frac{h}{\tilde{c} + h}}{\ln \rho} \right\rfloor,$$

where

$$\tilde{c} = \sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda} \frac{1 - \rho}{(1 - \rho_i)(1 - \rho_{i-1})} \right) b_i.$$

The optimal cost is then given by:

$$g_{sp} = \tilde{c} \frac{\rho^{\tilde{z}+1}}{1 - \rho} + h\left[\tilde{z} - \frac{\rho}{1 - \rho}(1 - \rho^{\tilde{z}})\right].$$

The proof can be found in Appendix 3.

3.4. The Multilevel Rationing Policy

Neither the FCFS nor the SP policy exploit the possibility of rationing the on-hand inventories. The SP policy should reduce average backorder costs by allocating production to the class with the highest backorder cost. One can intuitively generalize this allocation rule when there is on-hand inventory. In fact,

when the on-hand inventory level is low, cost may be reduced by backlogging classes with low unit backorder costs to reserve the available stock for future expensive demands. In that case, production is still allocated to on-hand inventory even though backorders exist. The Multilevel Rationing (ML) policy can reserve the inventory for future demands by rationing. We first define the ML policy formally. A more intuitive presentation follows.

An ML policy is characterized by n stock levels $z_1 \leq \cdots \leq z_n$. To be consistent with our notations, we take $z_0 = 0$. The controls are then:

(1) Production—Produce if and only if $x_0 < z_n$ or backlogs exist.

(2) Allocation—

- Production—Allocate the item to class k if and only if $x_0 = z_{k-1}$, and $m(\mathbf{x}) = k$. Otherwise, put the item into on-hand inventory.
- Stocks—An arriving demand of class i is satisfied with the stock if the inventory level is strictly above z_{i-1} ($x_0 > z_{i-1}$). It is back-ordered elsewhere ($x_0 \leq z_{i-1}$).

Note that no class k backorder is present in the system if $x_0 > x_{k-1}$. Note also that if all the z_k are different, the production-allocation rule can be restated as: If $x_0 = z_{k-1}$ and $x_k > 0$, then allocate to class k; allocate to on-hand inventory otherwise.

An alternative description of ML policies can be presented if the inventory is viewed to be composed of n (conceptual) *inventory layers.* Each inventory layer corresponds to a particular interval of on-hand inventory. More specifically, layer L_k corresponds to $z_{k-1} < x_0 \leq z_k$. With this definition, L_{k+1} is stacked on L_k, and each layer can contain a maximum number of parts equal to $z_1, z_2 - z_1, \ldots, z_n - z_{n-1}$ (so that the total physical capacity of the stock is equal to z_n). The current layer $L(t)$ then gives the layer corresponding to the current inventory position, $x_0(t)$. For instance, if $z_{k-1} < x_0(t) \leq z_k$, then $L(t) = L_k$. Figure 1 depicts an example of this structure for three classes of clients. Starting from a system at its base-stock level (inventory level $x_0 = z_n$), demands are first satisfied with parts coming from layer L_n. As soon as L_n is empty ($L(t)$ becomes L_{n-1}), they are satisfied with the next layer L_{n-1} until it empties, and so on. Note, however, that layer L_{n-1} is strictly reserved to classes 1, 2, . . . , $n - 1$, and class n demands cannot be satisfied from L_{n-1}. When $z_k > z_{k-1}$, if $L(t) = L_k$, demands belonging to classes 1, 2, . . . , k are satisfied from the stock and the other classes are back-ordered. As for production allocation, when a part is added to the stock, the on-hand inventory level increases so that L_k is refilled before L_{k+1}. Once again, if $L(t) = L_k$, there may be backorders of classes $k + 1, k + 2, \ldots, n$. As production continues and $L(t)$ becomes L_{k+1}, backorders of class $k + 1$ will be satisfied, while backorders of classes $k + 2, k + 3, \ldots, n$ continue to wait.

To give an example of how the ML policy functions, let us consider the three-class example of Figure 1. In this case the system starts with z_3 parts in the inventory ($L(0) = L_3$). As long as $L(t)$ equals L_3, all arriving demands, regardless of their class, are satisfied with the stock (like the FCFS and SP policies). When the current inventory level falls to z_2 (so that $L(t) = L_2$), the arriving demands of Class 3 are backordered. If the inventory level continues to decrease and reaches z_1 ($L(t) = L_1$), the demands of Class 2 are back-ordered, and so on. Hence, L_2 can only be used to satisfy demands of Classes 1 and 2. When a part is completed, if the stock is empty, it is assigned to satisfy a waiting demand of Class 1 (like the SP policy). But when all these demands are satisfied ($x_1 = 0$), the system produces to fill the layer L_1. It is only when L_1 is full (i.e., when $x_0 = z_1$) that the system produces to satisfy backordered demands of Class 2, and so on.

If the levels are not distinct, for instance if $z_{k-1} = z_k$, then when L_{k-1} is full, backordered demands of class k are satisfied before backordered demands of class $k + 1$. Then, when $x_k = x_{k+1} = 0$, the system produces to increase the inventory level. Thus, if $z_1 = \cdots = z_{n-1} = 0$, the ML policy is equivalent to the SP policy with $z = z_n$.

Note also that the recurrent states of the system are such that the on-hand inventory is less than z_n ($x_0 \leq z_n$) and such that if there are backorders ($\Sigma_{i=1}^{n} x_i > 0$), then the number of the most expensive waiting class is less than the number of the current layer ($m(\mathbf{x}) \leq k$, where k is such that $z_{k-1} < x_0 < z_k$). It follows that

the ML policy allows having a nonempty inventory with backordered demands.

An ML policy allocates both inventory and production, taking into account the current inventory and backorder positions. One would expect, then, that when its parameters are optimized, it should improve the performance of the system compared to the optimal FCFS and SP policies. Indeed, under certain assumptions it can be shown that the optimal ML policy is also optimal among all policies and solves the Minimization Problem (1) (de Véricourt et al. 2000b). In the following, the optimal parameters of the ML policy are presented as well as the optimal cost.

PROPERTY 3. *Construct the sequences z_k and g_k as follows:*

$$z_0 = g_0 = b_{n+1} = 0,$$

$$z_k - z_{k-1} = \left\lfloor \frac{\ln \dfrac{\rho_k(h + b_{k+1})}{\rho_k(h + b_k) + (1 - \rho_k)(g_{k-1} - (h + b_k)z_k)}}{\ln \rho_k} \right\rfloor$$

$$g_k = \left(z_k - \frac{\rho_k}{1 - \rho_k}\right)(h + b_{k+1}) + \left(g_{k-1} - \left(z_{k-1} - \frac{\rho_k}{1 - \rho_k}\right)(h + b_k)\right)\rho_k^{z_k - z_{k-1}}.$$

The optimal levels z_k^* and the optimal cost g_{ml} of the ML policy are equal to z_k and g_n.

The proof can be found in de Véricourt et al. (2000b).

3.5. Simple Insights

Based on the descriptions of the policies and the values of their respective optimal parameters and costs above, Property 4 summarizes the relationship in terms of optimal cost between the different policies:

PROPERTY 4. *Consider the costs g_{fcfs}, g_{sp}, and g_{ml} of the respective FCFS, SP, and ML optimal policies. We have:*

(1) $g_m \leq g_{sp} \leq g_{fcfs}$;

(2) $g_{fcfs} = g_{sp}$ if and only if the demand classes have identical b_k's;

(3) $g_{sp} = g_{ml}$ if and only if $z_1 = \cdots = z_{n-1} = 0$;

(4) If the demand classes are identical in b_k, then $g_{fcfs} = g_{sp} = g_{ml}$.

The proof can be found in Appendix 4.

Property 4 confirms our intuition that optimal cost performance of the policies improve with the degree of bias that can be offered to more expensive customers. It also follows from the property that when customers have almost identical backorder costs, the performance of the policies converge. In fact, if these costs are equal, all policies are identical. Furthermore, even if the backorder costs are not identical, there are cases in which the optimal SP policy may perform as well as the best ML policy. However, it also follows that if the optimal ML policy rations stocks, its performance must be superior to the other two policies. A complete investigation of these points will be undertaken in the next section to generalize and clarify some of these initial insights.

4. The Benefits of Effective Stock Allocation

In this section, we quantify the benefits of production and stock allocation by a numerical investigation to gain insights into the impacts of system parameters. To quantify these benefits we compare the optimal performances of the three control policies introduced earlier. The motivation for this investigation is twofold. On one hand, we would like to determine the benefits obtained by using an allocation policy that takes decisions based on the actual inventory and backorder positions. On the other hand, we would like to identify the situations in which simpler policies (that are described by less parameters) provide close-to-optimal performances.

Because Properties 1 through 3 are not constrained by the number of classes, the comparisons can, in principle, be performed for any number of classes of clients. For the sake of clarity, we first report the results of comparisons performed for the case of two classes of customers in §4.1. Later, in §4.2 we present generalizations and a discussion for multiple customer classes. Note that because ML policies are optimal, this comparison also provides the relative performances of the first two policies with respect to the optimal policy.

4.1. Systems with Two Customer Classes

To clarify the impact of different parameters of the system we study the following relative differences:

$$\Delta_{sp} = \frac{g_{sp} - g_{ml}}{g_{sp}} \quad \text{and} \quad \Delta_{fcfs} = \frac{g_{fcfs} - g_{ml}}{g_{fcfs}}.$$

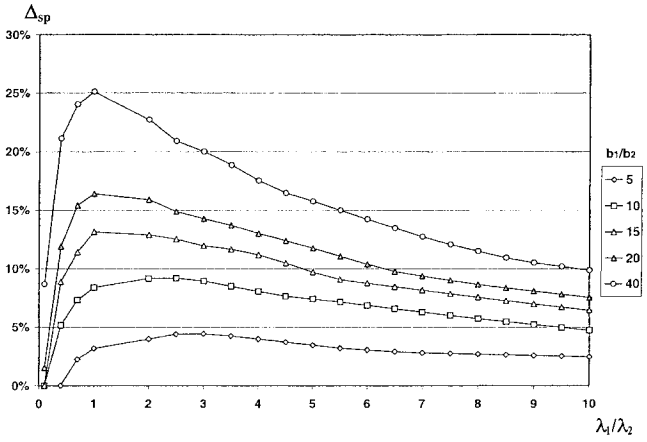
Δ_{fcfs} represents the relative benefit of implementing the optimal ML policy compared to implementing the optimal FCFS policy. Δ_{sp} represents the relative benefit of the optimal ML policy compared to the optimal SP policy. Δ_{fcfs} can then be interpreted as the relative gain when the optimal allocation policy is used in comparison to an optimal base-stock policy without any effort for rationing or production allocation. In addition, Δ_{sp} can be interpreted as the relative gain because of supplementing optimal production allocation by stock rationing.

More precisely, our investigation focuses on three important parameters: the utilization rate ρ, the relative backlog cost b_1/b_2, and the relative arrival rate λ_1/λ_2. We vary one of these quantities while keeping the others fixed. It is also useful to define h', the relative holding cost: $h' = h\rho/(\rho_1 b_1 + \rho_2 b_2)$. This quantity expresses the relative importance of the holding cost compared to the backlog costs. Unless otherwise indicated, we set h' equal to 0.01. Finally, we fix $\mu = 1$. All the other parameters of the system (λ_1, λ_2, b_1, b_2, h) can then be derived from the utilization rate, the relative arrival rate, the relative backlog cost, and the relative holding cost.

The expressions described in §3 were used to compute Δ_{sp} and Δ_{fcfs}. The different results obtained are plotted in Figures 2 through 6. A feature that is common to all figures is that the relative benefit of the optimal ML policy increases in b_1/b_2. This confirms our intuition because the ML policy reserves on-hand inventory for future expensive demands. For values of b_1/b_2 close to one, the optimal ML policy is equivalent to an SP policy. The cost reduction is significant when the ML policy is compared to the optimal FCFS policy (for instance, the relative difference is over 10% and may even reach up to 37% when $\lambda_1 = \lambda_2$ in Figure 3).

When λ_1/λ_2 is close to zero, demands of Class 1 are rare, so that the system behaves as if it were a single customer-class system with demands of Class 2, regardless of the policy which drives the system. The consequence is that priority allocation does not bring significant benefits in that case. The same argument holds when λ_1/λ_2 is large so that system is almost a single customer-class system with demands of Class 1. Hence, the effect of λ_1/λ_2 on the benefit of implementing an ML policy is nonmonotone. In fact, there exists a value of the ratio such that this benefit is maximum. This value is close to one ($\lambda_1 = \lambda_2$) when the ML policy is compared to the FCFS policy. But for the SP policy, this value can be larger. Remark that when $b_1 = b_2$ and $\lambda_1 = \lambda_2$, the two policies are the same. λ_1 must be larger than λ_2, such that rationing is required and differences can be observed between the two policies. Nevertheless, rationing the inventory can greatly (up to 25%) improve the performance of the system when the customers do not have identical backorder costs.

Figure 2 **Effect of λ_1/λ_2 on Δ_{sp} for Different Values of b_1/b_2, and with $\rho = 0.7$**



To summarize, stock rationing is especially beneficial for environments where the demand rates of customers with high backorder costs is of the same order as the demand rates of customers with low backorder costs and where the difference in backorder costs is significant.

Figures 4 and 5 depict the effects of ρ on the cost performance. The global effect of ρ on the cost performance can be nonmonotone (as seen in Figure 4). When ρ is small, the system has enough excess ca-

Figure 3 **Effect of λ_1/λ_2 on Δ_{fcfs} for Different Values of b_1/b_2 and with $\rho = 0.7$**

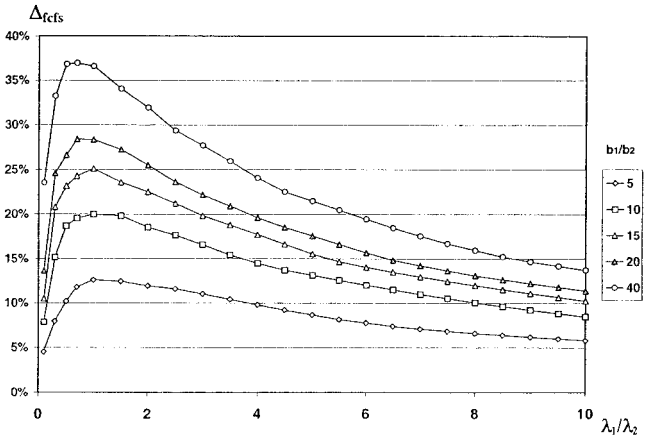
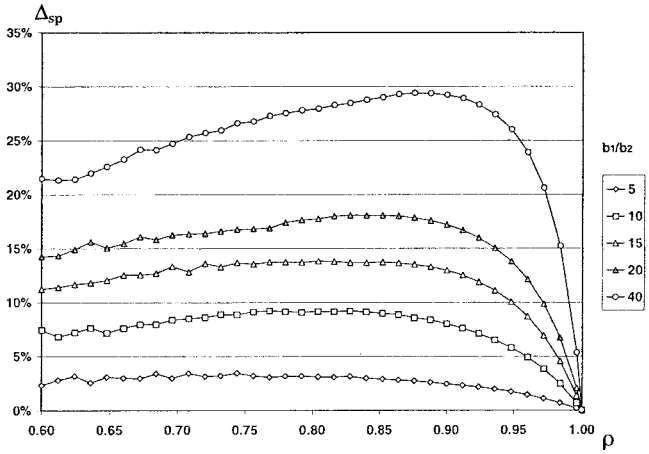


Figure 4 **Effect of ρ on Δ_{sp} for Different Values of b_1/b_2 and with $\lambda_1 = \lambda_2$**



pacity to satisfy the arriving demands, and no rationing is required (the small variations in cost that appear in Figures 4 and 5 when $\rho \simeq 0.6$ are because of the discrete nature of the problem). When ρ is large, stockouts become more frequent and the policies differ only when there are significant backorders. Hence, the ML policy and the SP policy are equivalent so that their performance is almost identical. Note, however, that for large ρ, the benefit of rationing can still be significant (for example, when $\rho = 0.9$, the relative difference can reach 30% in Figure 4). Furthermore, as ρ approaches one, the relative difference Δ_{fcfs} seems to approach a finite limit (see in Figure 5). This limit can be interpreted as the maximum benefit that can be expected by implementing a priority discipline compared to a FCFS discipline. These limit arguments are difficult to state precisely from the numerical investigation and will be proven in the following section through a heavy-traffic analysis.

The final experiment investigates the effects of the relative holding cost, h', on the performance. This effect was seen to be nonmonotone in the corresponding lost-sales model (see Ha 1997b). The results summarized in Figure 6 indicate that, in the backorder case, this effect is monotone. This highlights an important difference between FCFS and ML policies in backorder environments. In the lost-sales case, if the optimal base-stock levels are low, there is little difference between the two policies. In the backorder-cost case, in contrast, even if the optimal base-stock levels are small, a significant performance difference remains between the two policies because of optimized allocation of production.

4.2. Generalizations to Systems with Multiple Customer Classes

The main difficulty in carrying out a numerical study with more than two customer classes is the number of parameters which have to be specified. However, some of the previous insights can be generalized by focusing on some of the key parameters.

A critical parameter in an n-class system is b_1/b_n, the ratio of the highest to the lowest backorder cost. When b_1/b_n is close to one, all the backorder costs are almost equal, so that there is little need for stock rationing or priority setting between classes. In this case, the performances of all the three policies are similar. At the other extreme, when the ratio b_1/b_n is very large, costs can be reduced by rationing and priority setting, at least between the first and the nth class.

For instance, Figures 2 through 6 reveal that, in the two-class system, as b_1/b_2 increases, Δ_{fcfs} and Δ_{sp} also increase. Based on the previous comments, we expect to see the same qualitative impact of b_1/b_n on Δ_{fcfs} and Δ_{sp} when the value of b_n is set to b_2 (of the two-class system) and the utilization rate of the system is held constant.

Figure 5 **Effect on ρ on Δ_{fcfs} for Different Values of b_1/b_2 and with $\lambda_1 = \lambda_2$**

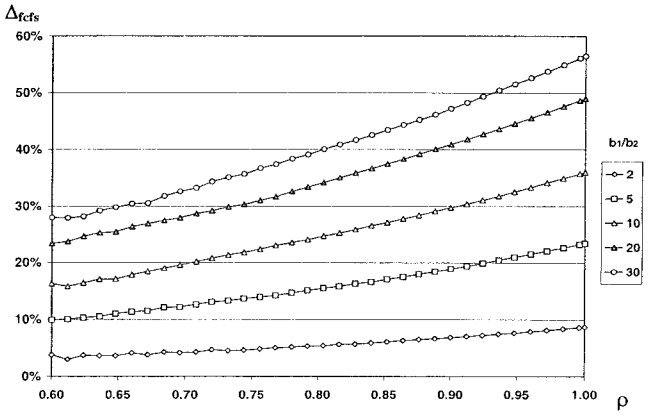
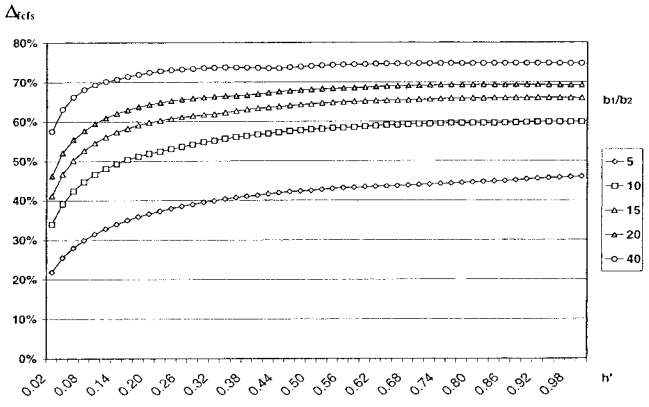


Figure 6 **Effect of h' on Δ_{fcfs} for Different Values of b_1/b_2 with $\rho = 0.9$, $\lambda_1 = \lambda_2$**



Furthermore, we can revisit Figures 4 and 5 in the case of n customer classes for a second observation. The figures show that Δ_{sp} approaches zero, while Δ_{fcfs} attains a positive value, when the utilization rate approaches one. We can expect that both of these results will hold for n customer classes for the same reasons as in the two-class system. In fact, this property will be formalized in the next section.

Note that introducing a new customer class to the system increases the costs regardless of the control policy employed. Let us briefly describe this impact for the ML policy. We consider a two-class system with identical arrival rates, and we investigate the cost increase because of the addition of a third customer class with arrival rate $\lambda_3 = \lambda_1 = \lambda_2$ and backorder cost b.

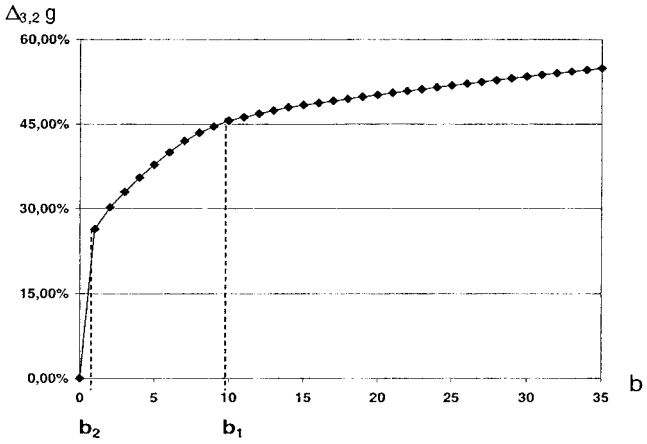
For instance, consider the following backorder costs for the two classes of the original system: $b_1 = 10$, $b_2 = 1$. Let us assume that $h = 1$, $\mu = 1$, and $\lambda_1 = \lambda_2 = 0.3$ (the corresponding utilization rate is equal to 0.6). Let g_2 be the average cost for this system when the ML policy is used. g_3 corresponds to the optimal average cost when the system satisfies a third demand class with arrival rate 0.3 (the utilization rate is then equal to 0.9). Remark that, with the introduction of a third part, the optimal ML policy uses three parameters.

Figure 7 depicts the impact of b on the relative cost increase $\Delta_{3,2}g = (g_3 - g_2)/g_3$ (when the other parameters are fixed). As expected, $\Delta_{3,2}g$ is a positive increasing function of b. In addition, $\Delta_{3,2}g$ has a concave shape. In particular, when $b < b_2$, the parameters of the optimal ML policy is such that the third class has the lowest allocation priority. The relative cost increases, nevertheless, in a sharp manner because additional inventory needs to be held for the new class and more stock reservation is required for the other two classes. When $b_2 < b < b_1$, the new class has the second priority and is protected against Class 2. The relative cost increase is still significant. Finally, when $b > b_1$, the new class is protected against both Classes 1 and 2. In fact, when b is much larger than the other backorder costs, its effect completely dominates the system, and $\Delta_{3,2}g$ approaches 100%.

5. Stock Allocation Under High Capacity Utilization

The results of §4 indicate that when the production capacity of the system is very close to the total demand rate, the system can exhibit particular properties. This heavy-traffic regime can be extremely relevant for certain industries such as the semiconductor industry, which operate very close to a capacity saturation level. Unfortunately, the properties that are particular to this regime are difficult to extract from numerical experiments. In this section, we complement the previous numerical results with an analytical study of the heavy-traffic regime, wherein the ca-

Figure 7 **Performance of the ML Policy for $n = 2$ Versus $n = 3$**



pacity of production becomes barely sufficient to satisfy arriving demands, that is, when ρ tends to one.

THEOREM 1. *Suppose that* $\lambda_k = a_k\rho$, *such that* $\Sigma_{k=1}^n a_k = 1$. *Then we have*

$$\lim_{\rho\to 1} \Delta_{sp} = 0 \tag{2}$$

$$\lim_{\rho\to 1} \Delta_{fcfs} = 1 - \frac{\ln s}{\ln r} \tag{3}$$

where $r = h/(h + \Sigma_{k=1}^n a_k b_k)$ *and* $s = h/(h + b_n)$.

The proof can be found in Appendix 5.

A simplified interpretation of Theorem 1 is that there is little benefit in optimized stock allocation when the system is operated at extremely high utilization rates (for instance, at 99% capacity utilization). For a more complete managerial interpretation, this interpretation has to be combined with the results from §4, which demonstrate that the convergence to the limiting value is rather slow. The final conclusion then, is the ML policy should be preferred even at high utilization levels (i.e., 95 to 98%) because it can result in substantial benefits. Only in very extreme cases do the relative benefits of dynamic allocation diminish, but even then consistency and robustness properties may favor ML policies.

Theorem 1 also implies that the maximum relative benefit achieved by the ML policy with respect to the FCFS policy is finite. Note that if $b_n \simeq \Sigma\, a_k b_k$ (which implies that the b_k are close), $r \simeq s$, and there is no benefit in implementing an ML policy. This is reminiscent of Property 4. On the other hand, if b_n is small compared to the other b_k's, the limit of the relative difference can go up to 100%.

6. Fill-Rate Constraints

In the preceding sections, the dissatisfaction of a waiting customer of class k was modeled by a linear cost rate b_k. An alternative approach that is frequently used in practice is to express this dissatisfaction through a service-level measure. One of the most commonly employed measures is the fill rate (see Nahmias and Demmy 1981 or Zipkin 2000): the proportion of items directly satisfied from stock. For instance, a fill-rate constraint of $1 - \alpha_k$ specifies that the fraction of demands of class k satisfied from the stock (without having to wait) must be higher than or equal to $1 - \alpha_k$.

Assume that the required fill-rate level, $1 - \alpha_k$, of each demand class is an exogenous parameter specified by a contractual agreement. We define f_k^π to be the effective fill rate, i.e., the fraction of arriving demands of class k not filled from the stock under the control policy π (we will also use the notation f_k when no confusion is possible). In this case, the α_k are given by the fill-rate requirements, and the manager must control the system to minimize the average holding cost, while ensuring that the effective fill rates, $(1 - f_k)$, satisfy the requirements.

When the clients have different fill-rate requirements, production and stock allocation policies that give priorities to certain classes should improve performance. Even though it is difficult to precisely characterize optimal policies in this scenario, the three policies discussed earlier are intuitively plausible and interesting.

Let us first qualitatively compare the FCFS and SP policies. Neither of these policies ration the on-hand stock. Indeed, their only difference is because of the respective production-allocation rules when demands are backordered. Note, however, that even though the allocation of production to backorders affects the average backorder times, it does not have any effect on

the respective effective fill rates. Hence, any nonidling production-allocation policy with the same base-stock level (in the absence of stock rationing) has equivalent holding-cost performance. Because of this equivalence, the analysis in this section will be restricted to the class of FCFS and ML policies.

In the following, we assume that the classes of demands are ordered such that $\alpha_1 < \cdots < \alpha_n$. The definition of an ML policy is then unchanged.

The optimal $\hat{z}$ of the FCFS policies under the fill-rate constraints is given by the following property:

PROPERTY 5. *The optimal FCFS policy is characterized by the base-stock level equal to*

$$\hat{z} = \left\lceil \frac{\ln(\alpha_1)}{\ln \rho} \right\rceil.$$

The optimal holding cost is then given by

$$g^{\alpha}_{fcfs} = h\left[\hat{z} - \frac{\rho}{1-\rho}(1-\rho^{\hat{z}})\right].$$

The proof can be found in Appendix 6.

Thus, all the effective fill rates $1 - f_k$ of the demand classes satisfy the constraint $1 - \alpha_1$, which is the most restrictive requirement. This is the main drawback of the FCFS policy: For a fixed α_1 (and a fixed base-stock level), regardless of the values of α_k, the cost and the effective fill rate stay unchanged. In other words, the performance of the system is determined uniquely by the most stringent requirement.

An ML policy, on the other hand, does not suffer from this drawback. Inventory-level-dependent allocation of the stock allows a flexibility that enables fitting the different effective fill rates to their respective constraints $1 - \alpha_k$.

PROPERTY 6. *Construct the sequences $z_k - z_{k-1}$, ρ_k and f_k starting from $k = n$, with $f_{n+1} = 1$, as follows:*

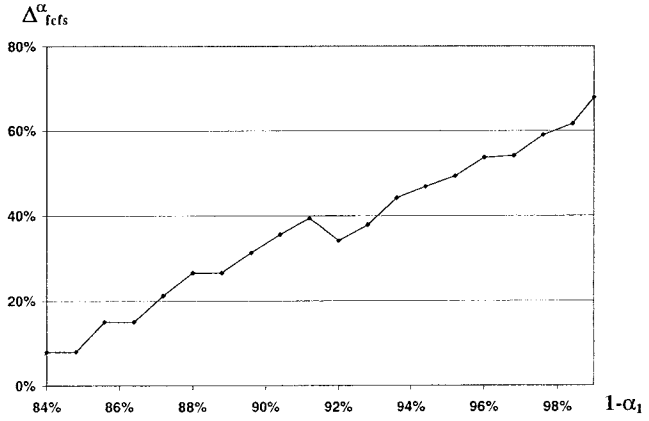
$$\rho_k = \frac{\sum_{i=1}^{k} \lambda_i}{\mu}, \qquad z_k - z_{k-1} = \left\lceil \frac{\ln \frac{\alpha_k}{f_{k+1}}}{\ln \rho_k} \right\rceil \qquad f_k = \rho_k^{z_k - z_{k-1}} f_{k+1}.$$

Then, we obtain

(1) the optimal levels of the ML policy are equal to z_k (with $z_0 = 0$);

(2) its optimal cost g^{α}_{ml} is equal to

Figure 8 **Effect of $1 - \alpha_1$ on Δ^{α}_{fcfs} with $\lambda_1 = \lambda_2$, $1 - \alpha_2 = 80\%$, and $\rho = 0.9$**



$$g^{\alpha}_{ml} = h \sum_{k=1}^{n} I_k f_{k+1}$$

where

$$I_k = z_k - z_{k-1}\rho_k^{z_k - z_{k-1}} - \frac{\rho_k}{(1-\rho_k)}(1 - \rho_k^{z_k - z_{k-1}}); \quad \text{and}$$

(3) $1 - f_k$ represents the fill rate of customer k under the optimal ML policy.

The proof can be found in Appendix 7.

We are now able to evaluate the benefit of implementing the ML policy compared to implementing the FCFS policy. We consider the relative difference $\Delta^{\alpha}_{fcfs} = (g^{\alpha}_{fcfs} - g^{\alpha}_{ml})/g^{\alpha}_{fcfs}$, and we take $n = 2$.

Figure 8 presents the evolution of Δ^{α}_{fcfs} for increasing fill-rate constraints for Class 1, with $\lambda_1 = \lambda_2$, $1 - \alpha_2 = 80\%$, and $\rho = 0.9$. For values of the service levels commonly used in practice (around 95%, the benefit of implementing an ML policy compared to the FCFS policy is very significant (60%). When $1 - \alpha_1$ is close to 80% (that is, when $1 - \alpha_1 \simeq 1 - \alpha_2$), both policies are equivalent ($z_1 = 0$), and Δ^{α}_{fcfs} tends to zero. When $1 - \alpha_1$ increases, the flexibility of the ML policy allows adjustment of the stocks for the respective fill-rate constraints, while the FCFS policy maintains all the achieved service levels at $1 - \alpha_1$. The higher $(1 - \alpha_1)$ is, the more valuable the impact due to this flexibility becomes. The relative cost difference then ap-

proaches to a finite limit, which is the maximum cost reduction that can be attained by implementing an ML policy.

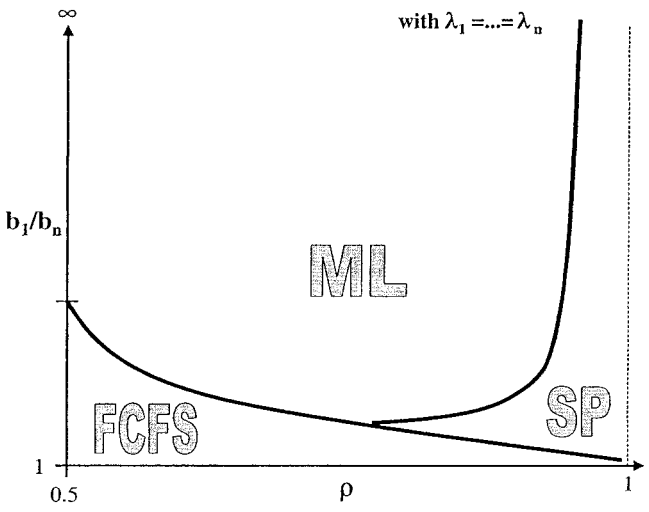
Differentiating classes of customers by their relative backorder costs may be a difficult task. In this sense, the service-level constraint approach provides an alternative framework for measuring the relative importance of customers. The end result is that inventory-level-dependent allocation policies can improve performance significantly (by up to 30% in Figure 8) even for a relatively small (10%) difference in service levels. To understand how this significant difference in performance is achieved, let us compare the effective fill rates and average number of backlogs for optimal FCFS and ML policies using the parameters of Figure 8 (with $1 - \alpha_1 = 90\%$ and $1 - \alpha_2 = 80\%$). The optimal FCFS policy has effective fill rates of 90.1% and average backlog levels of 0.44 for both classes. The optimal ML policy, in contrast, has effective fill rates of 91.6% and 81.5% for Classes 1 and 2, respectively. The corresponding average backlog levels are 0.068 and 1.52 for Classes 1 and 2, respectively, (the computation of average backlog levels for the MP policy is based on the proof of Property 3, see de Véricourt et al. 2000b). As expected, when inventory-level-dependent allocation is employed, the bias towards stringent customers in terms of required fill rates translates into a bias in terms of average backlog levels (or equivalently, into average waiting times in the backlog queue).

Two results emerge from the analysis in this section. Firstly, when fill rates are the appropriate measure, service-level differentiation cannot be achieved without stock rationing. Secondly, when dynamic stock rationing is performed optimally, important benefits can be obtained even for small differences in service levels of customers.

7. Conclusion

We have presented three control policies for a make-to-stock production system with multiple classes of customers. Our initial contribution is to obtain the optimal parameters of ML policies for a fill-rate constraint formulation and of FCFS and SP policies for both fill rate and backorder cost formulations. We then studied and compared the optimal performances of the three policies to shed light onto the potential benefits of stock allocation.

Figure 9 When Are Other Policies as Good as ML Policies?



The ML policy always outperforms the other two policies. This robustness is the first evident advantage of the ML policy. A second pertinent issue is the following one: Are there systems for which other base-stock-type policies (which use fewer parameters) are almost as good as ML policies? Figure 9, which summarizes the numerical investigation, indicates that there are parameter values for which the performance of other policies is comparable to the performance of ML policies. Nevertheless, the figure also indicates that for an important range of parameters the ML policy is the best choice. Furthermore, because the optimization of an ML policy is fairly easy, once the inventory-tracking system that enables the implementation of such a policy is available, it is not too difficult to readjust the policy parameters to cope with changing system parameters.

An important last question is how do the qualitative results (such as Figure 9) depend on the modeling assumptions? The assumptions of Poisson arrivals and exponential service times significantly facilitate exact computation and precise statements on optimal control. Incorporation of more general arrival or service processes in the model can be expected to

modify the quantitative results. Fortunately, other results in make-to-stock queues indicate that the qualitative insights that are obtained in the simple Markovian framework are quite robust. In particular, Ha (2000) provides evidence in this direction for a stock-rationing problem with nonexponential service times and lost sales. The actual optimal policy is more complicated than a (lost-sales) ML policy in this case, but the ML policy results in performances that are remarkably close to optimal.

Acknowledgments

The authors would like to thank the referees, the senior editor, and the editor for several suggestions that have led to improvements in the presentation of the paper.

Appendix

Appendix 1.

Formulation of the Optimal Control Problem. A control policy states the action to take at any time given the current state $\mathbf{x}(t)$. The investigation can be restricted to Markovian policies because the optimal policy belongs to this class. Let $C^{\pi}(\mathbf{x}) = (C^{\pi}_{\emptyset}(\mathbf{x}), \ldots, C^{\pi}_{n}(\mathbf{x}))$ the control associated with a policy π defined by:

$$C^{\pi}_{\emptyset}(\mathbf{x}) = \begin{cases} -1 & \text{not to produce} \\ k \quad 1 \le k \le n, & \text{to allocate production to backorders of class } k, \end{cases}$$

$$C^{\pi}_{k}(\mathbf{x}) = \begin{cases} 0 & \text{to satisfy an arriving class } k \text{ demand from the inventory} \\ k & \text{to back-order an arriving class } k \text{ demand,} \quad k \ge 1. \end{cases}$$

$C^{\pi}_{\emptyset}$ corresponds to the control of the production of the facility. When there are backorders, it also states which class of customers has to be satisfied. C^{π}_{k} for $k \ge 1$ corresponds to the rationing of class k. By g^* we denote the optimal average cost.

Let $\mathbf{e}_i$ be the unit vector of dimension i. Without loss of generality, we can uniformize transition rates by taking $\Sigma_{i=1}^{n} \lambda_i + \mu = 1$. The value function v^* for the corresponding Markov decision problem can be shown to satisfy the following optimality equations:

$$v^*(\mathbf{x}) + g^* = c(\mathbf{x}) + \mu T_0 v^*(\mathbf{x}) + \sum_{k=1}^{n} \lambda_k T_k v^*(\mathbf{x}), \tag{A1}$$

where the operators T_k are

$$T_0 v(\mathbf{x}) = \min_{1 \le i \le n} [v(\mathbf{x}), v(\mathbf{x} + \mathbf{e}_0), v(\mathbf{x} - \mathbf{e}_i \mathbf{I}_{\{x_i < 0\}})]$$

where $\mathbf{I}$ is the indicator function

$$T_k v(\mathbf{x}) = \min[v(\mathbf{x} + \mathbf{e}_k), v(\mathbf{x} - \mathbf{e}_0)], \quad \text{for } k \text{ such that } 1 \le k \le n.$$

By standard results in dynamic programming, the optimal policy can then be obtained through the optimality equation of the above Markov decision process. This control problem has been previously described by Ha (1997c) in the case of two demand classes and has been generalized to the multicustomer-class case in de Véricourt et al. (2000b).

Appendix 2.

Proof of Property 1. Consider a FCFS policy with the base-stock level z. The recurrent states are such that $x_0 \times \Sigma_{i=1}^{n} x_i = 0$ with $x_0 \le z$. For these states, the random variable that equals $z - X_0$ if $X_0 > 0$, and that equals $\Sigma_{i=1}^{n} X_i + z$ elsewhere, is equivalent to the length of an M/M/1 queue. It follows that $P(X_0 = i) = (1 - \rho)\rho^{z-i}$, $i = 1, \ldots, z$, and $P(X_0 = 0) = \rho^{z}$.

Furthermore, given that the on-hand inventory is empty, the system is equivalent to a FCFS multiclass M/M/1 queue. It follows that for $i > 0$, $E[X_i \mid \{X_0 = 0\}] = \hat{p}_i/(1 - \hat{p}_i)$, where $\hat{p}_i = \lambda_i/(\mu - \Sigma_{j=1, j \ne i}^{n} \lambda_j)$ (see Buzacott and Shanthikumar 1993). The cost $g_{fcfs}(z)$ of the FCFS policy with the base stock z is then equal to

$$\begin{aligned} g_{fcfs}(z) &= \sum_{k=1}^{n} \frac{\hat{\rho}_k b_k}{1 - \hat{\rho}_k} \rho^{\hat{z}} + h\left[z - \frac{\rho}{1 - \rho}(1 - \rho^{z})\right] \\ &= \hat{b}\frac{\rho^{\hat{z}+1}}{1 - \rho} + h\left[z - \frac{\rho}{1 - \rho}(1 - \rho^{z})\right]. \end{aligned} \tag{A2}$$

Taking the first difference of $g_{fcfs}(z)$ in z we obtain

$$\Delta g_{fcfs}(z) = g_{fcfs}(z + 1) - g_{fcfs}(z) = h - (\hat{b} + h)\rho^{z+1},$$

which is nondecreasing in z. The minimum is reached at $\hat{z} = \min_z \{\Delta g_{fcfs}(z) > 0\}$ leading to

$$\hat{z} = \left\lceil \frac{\ln \dfrac{h}{\hat{b} + h}}{\ln \rho} \right\rceil,$$

and replacing $\hat{z}$ in (5) we obtain g_{fcfs}. □

Appendix 3.

Proof of Property 2. The proof is very similar to the FCFS case. Consider an SP with the base stock z. The distribution of X_0 is $(1 - \rho)\rho^{z-i}$, $i = 1, \ldots, z$. Given that X_0 is empty, the system is the same as a multiclass queue with preemptive priority. The average number of backordered demands of class i is then equal to (see Gross and Harris 1985), $\rho_i/(1 - \rho_i) - \rho_{i-1}/(1 - \rho_{i-1})$, and a straightforward computation leads to

$$g_{sp}(z) = \tilde{c}\frac{\rho^{z+1}}{1 - \rho} + h\left[z - \frac{\rho}{1 - \rho}(1 - \rho^{z})\right].$$

Taking the first difference of $g_{fcfs}(z)$, which is nondecreasing in z, we obtain the optimal base-stock level. The optimal average cost follows by a direct computation. □

Appendix 4.

Proof of Property 4. An ML policy is optimal for Problem (1) (see de Véricourt et al. 2000b) leading to $g_{ml} \le g_{fcfs}$ and $g_{ml} \le g_{sp}$. Furthermore, $\rho/(1-\rho)\hat{b}$ and $\rho/(1-\rho)\tilde{c}$ are the average costs of the corresponding multiclass queue with respectively FCFS and preemptive priority (see the proofs of properties 1 and 2). Because a $c\mu$ rule is optimal for this problem (see Baras et al. 1985), $\tilde{c} \le \hat{b}$. Thus, we have $\hat{z} \ge \tilde{z}$ and $g_{sp} \le g_{fcfs}$.

Note then that the optimal costs of the FCFS and SP policies are equivalent to the optimal cost of the well-known single-part-type, single-server problem (see, for instance, Veatch and Wein 1996) with the arrival rate λ, the service rate of μ, the holding cost h, and a backlog cost equal to $\hat{b}$ for the FCFS policy and $\tilde{c}$ for the SP policy. Hence, $\tilde{c} \le \hat{b}$ leads to $g_{sp} \le g_{fcfs}$, proving the first part of the proposition.

It also follows that $g_{sp} = g_{fcfs}$ if and only if $\tilde{c} = \hat{b}$, which can be shown to be equivalent to the equality of all the b_k, giving us the second part of the property.

The third part comes directly from the definition of the ML policies. Furthermore, if all the b_k are the same, then $z_1 = \cdots = z_{n-1} = 0$. Hence, from the third part of the property the last one is also true. □

Appendix 5.

Proof of Theorem 1. Using Property 1 we obtain for the optimal FCFS policy,

$$\hat{b} = \sum a_k b_k \qquad \frac{\ln r}{\ln \rho} - 1 < \hat{z} \le \frac{\ln r}{\ln \rho}$$

$$r \le \rho^{\hat{z}} < \frac{r}{\rho}$$

$$\frac{\sum a_k b_k}{1-\rho} r + h\left[\frac{\ln r}{\ln \rho} - 1 - \frac{\rho(1-r)}{1-\rho}\right] \le g_{fcfs} \le \frac{\sum a_k b_k}{1-\rho}\frac{r}{\rho} + h\left[\frac{\ln r}{\ln \rho} - \frac{\rho - r}{1-\rho}\right]. \tag{A3}$$

Hence, we have from (A3)

$$-\frac{h \ln(r)}{1-\rho} + cst_{fcfs}^{inf} + o(1-\rho) \le g_{fcfs} \le -\frac{h \ln(r)}{1-\rho} + cst_{fcfs}^{sup} + o(1-\rho), \tag{A4}$$

where $cst_{inf}^{fcfs} \le cst_{sup}^{fcfs}$ are two constants.

Following the same steps we obtain similar results for the SP and ML cases,

$$-\frac{h \ln(s)}{1-\rho} + cst_{sp}^{inf} + o(1-\rho) \le g_{sp} \le -\frac{h \ln(s)}{1-\rho} + cst_{sp}^{sup} + o(1-\rho), \tag{A5}$$

$$-\frac{h \ln(s)}{1-\rho} + cst_{ml}^{inf} + o(1-\rho) \le g_{ml} \le -\frac{h \ln(s)}{1-\rho} + cst_{ml}^{sup} + o(1-\rho), \tag{A6}$$

where cst_{sp}^{inf}, cst_{sp}^{sup}, cst_{ml}^{inf}, and cst_{ml}^{sup} are four constants. Equations (A4), (A5), and (A6) give us the desired result. □

Appendix 6.

Proof of Property 5. For the optimal FCFS policy, we have $f_1 = \cdots = f_n = P(X_0 = 0)$. It follows that the optimal base-stock level is given by $\hat{z} = \min_z\{P(X_0 = 0) \le \min_k \alpha_k\}$ with $P(X_0 = 0) = \rho^{\hat{z}}$. Hence, we obtain $\hat{z} = \lceil \ln(\alpha_1)/\ln \rho \rceil$, and the optimal average cost can be derived directly. □

Appendix 7.

Proof of Property 6. Consider a given ML policy with its n stock levels $z_1, \ldots, z_n$ (and taking $z_0 = 0$). Under this policy the probability distribution of the on-hand inventory X_0 is equal to, for $x \in L_k$ $(z_{k-1} < x \le z_k)$,

$$P(X_0 = x) = \prod_{i=k+1}^{n} \rho_i^{z_i - z_{i-1}}(1-\rho_k)\rho_k^{z_k - x}. \tag{A7}$$

Furthermore, $f_k = P(X_0 \le z_k) = 1 - P(X_0 > z_k)$, and using Equation (A7) a straightforward calculation leads to

$$f_k = \prod_{i=k}^{n} \rho_i^{z_i - z_{i-1}}. \tag{A8}$$

To compute the optimal ML policy, we minimize the differences $z_k - z_{k-1}$ such that $f_k \le \alpha_k$. Hence, for $0 < k \le n$, Equation (A8) leads to $f_k = e\rho_k^{z_k - z_{k-1}} f_{k+1}$ (with $f_{n+1} = 1$), and $z_k - z_{k-1} = \min_{\Delta z}\{\rho_k^{\Delta z} f_{k+1} \le \alpha_k\}$ gives us $z_k - z_{k-1} = \lceil \ln(\alpha_{k-1}/f_{k+1})/\ln \rho_k \rceil$. The holding cost can then be obtained using Equation (A7). □

References

Baras, J. S., A. J. Dorsey, M. Makowski. 1985. Two competing queues with linear costs: The μc rule is often optimal. *Adv. Appl. Probab.* **17** 186–209.

Buzacott, J. A., J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Cohen, M. A., P. R. Kleindorfer, H.L. Lee. 1988. Service constrained (s,S) inventory systems with priority demand classes and lost sales. *Management Sci.* **34** 482–499.

Frank, K. C., R. C. Zhang, I. Duenyas. 1999. Inventory control and rationing in a system with deterministic and stochastic sources of demand. Working paper.

Gross, D., C. M. Harris. 1985. *Fundamentals of Queueing Theory.* Wiley, New York.

Ha, A. 1997a. Optimal dynamic scheduling policy for a make-to-stock production system. *Oper. Res.* **45**(1) 42–53.

——. 1997b. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Sci.* **43**(8) 1093–1103.

——. 1997c. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Res. Logistics* **44** 458–72.

——. 2000. Stock-rationing in an $M/E_k/1$ make-to-stock queue. *Management Sci.* **46**(1) 77–87.

Jackson, P. 1988. Stock allocation in a two-echelon distribution system or 'what to do until your ship comes in'. *Management Sci.* **34**(7) 880–895.

Mc Gavin, E. J., L. B. Schwarz, J. E. Ward. 1993. Two-interval inventory-allocation policies in a one-warehouse *N*-identical-retailer distribution system. *Management Sci.* **39**(9) 1092–1107.

Nahmias, S., W. S. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Sci.* **27**(11) 1236–1245.

Peña-Perez, A., P. Zipkin. 1997. Dynamic scheduling rules for a multiproduct make-to-stock queue. *Oper. Res.* **45**(6) 919–930.

Topkis, D. M. 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with *n* demand classes. *Management Sci.* **15**(3) 160–176.

Veatch, M., L. M. Wein. 1996. Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. Res.* **44**(4) 634–647.

de Véricourt, F., F. Karaesmen, Y. Dallery. 2000a. Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies. *Oper. Res.* **48**(5) 811–819.

——, ——, ——. 2000b. Optimal stock rationing for a capacitated make-to-stock production system. Tech. Rep. École Centrale, Paris, France.

Wein, L. M. 1992. Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.* **40**(4) 724–735.

Zipkin, P. H. 2000. *Foundations of Inventory Management.* McGraw-Hill, New York.

The consulting Senior Editor for this manuscript was John Buzacott. This manuscript was received December 31, 1999, and was with the authors 357 days for 6 revisions. The average review cycle time was 28 days.