The Impact of Adding a Make-to-Order Item to a Make-to-Stock Production System

Awi Federgruen • Ziv Katalan
Graduate School of Business, Columbia University, New York, New York 10027
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Stochastic Economic Lot Scheduling Problems (ELSPs) involve settings where several items need to be produced in a common facility with limited capacity, under significant uncertainty regarding demands, unit production times, setup times, or combinations thereof. We consider systems where some products are made-to-stock while another product line is made-to-order. We present a rich and effective class of strategies for which a variety of cost and performance measures can be evaluated and optimized efficiently by analytical methods. These include inventory level and waiting-time distributions, as well as average setup, holding, and backlogging costs. We also characterize how strategy choices are affected by the system parameters. The availability of efficient analytical evaluation and optimization methods permits us to address the impact of product line diversification or standardization on the performance of the manufacturing system, in particular the logistical implications of adding low-volume specialized models to a given make-to-stock product line.
(*Performance/Productivity Manufacturing*; *Queues*; *Priorities*; *Inventory/Production*; *Multi-Item*)

1. Introduction and Summary

Stochastic Economic Lot Scheduling Problems (ELSPs) involve settings where N items need to be produced in a common facility with limited capacity, under significant uncertainty regarding demands, or unit production and setup times. Examples include a focused factory or work center of limited capacity that is dedicated to a group of items.

Stimulated by the current Just-in-Time and Quick Response movements, managers increasingly advocate Make-to-Order (MTO) systems supported by zero or small inventories, agile enough to guarantee short response times. To enhance Quick Response, buyers of custom products are increasingly encouraged to order significantly in advance of their actual needs. Various information sharing procedures within companies and between companies in a supply chain, e.g., via Electronic Data Interchange or private satellite communication systems, provide mechanisms to convey orders rapidly, which results in additional advanced warning. In parallel, the trend toward increased product variety generates the necessity to produce certain items to customer specifications, i.e., to order. Thus MTO is now often an option or a necessity for certain items. In the presence of the (MTO) option, considerable confusion prevails about how to exercise it. As stated in Perkins (1994):

> The gospel of Just-in-Time says stock nothing. The standard textbook approach says stock everything, particularly the items for which demand is most unpredictable. And flexible manufacturing equipment salesfolk say stop obsessing about stock, just buy fancier machines for your company.

Make-to-Stock (MTS) systems have traditionally been viewed as entirely distinct and incompatible with the MTO philosophy. However, some have come to realize that a system with a diverse product line and customer base can only be, or alternatively, is best served with an appropriate combination of these two extreme philosophies.

In this paper, we present for hybrid systems a

0025-1909/99/4507/0980$05.00

complete spectrum of options on how to prioritize the MTO item vis-à-vis the production of the MTS items. Next, we show how for each of these options, a variety of performance measures can be evaluated and optimized efficiently by analytical methods (e.g., inventory and waiting time distributions, and average setup, holding, and backlogging costs). Third, we show how various performance measures become stochastically larger or smaller as one moves along the spectrum of priority options. Fourth, we derive stability conditions and expressions for the effective utilization rate under a variety of the above options. Finally, we report on an extensive numerical study conducted to investigate some of the main choices presented in this paper. In particular, we have gauged which of the priority options is to be preferred, and under what circumstances. The numerical study also addresses the impact of product-line diversification or standardization on the performance of the manufacturing system, in particular the logistical implications of adding low-volume specialized models to a given product line. These include the allocation mechanism of the production capacity among MTO and MTS items and the degree to which additional MTO items result in increased inventory requirements for the MTS items.

In this paper we consider a single MTO item. (See Federgruen and Katalan 1995a for extensions to settings where several distinct MTO items need to be distinguished.) The MTO item is not primarily distinguished by the fact that it is supported by no or just a small inventory, but rather by the type of priority it is given in the overall production strategy. As mentioned, often the item cannot be stocked since it must be made to customer specifications. We thus, henceforth, refer to this item as the B-item, and to all others as the A-items. A production/inventory strategy for a hybrid system consists of two components: (i) an appropriate *interruption* discipline determining when to switch from the A-items to the B-item; (ii) a *production* strategy that determines which and how much of the A-items to produce in the absence of interruptions for the B-item. If the B-item is purely MTO, the facility continues producing the B-item until all its demand is satisfied. More generally, production continues until a given (small) inventory is built up.

Several *interruption* disciplines (i) may be considered: Under *preemptive* priority rules, the facility switches to the B-class as soon as an order for a B-item arises, but under *nonpreemptive* priority rules only at an arbitrary production completion epoch of an A-item, or only at some or all completion epochs at which the production strategy in (ii) calls for a switch-over between distinct A-items.

As far as the *production* strategy is concerned, to design overall strategies that are analytically tractable, we consider the class of base-stock policies with a general periodic sequence. Under such a policy we continue production of a given item until a specific target inventory level is reached; the different items are produced in a fixed periodic sequence that repeats itself after every M ($\geq N$, say) items. For a given periodic sequence, the analytical method of Federgruen and Katalan (1994) enables the efficient determination of optimal base-stock levels and the evaluation of all relevant performance measures. For systems with up to 50 items (say), this can be done in just a few seconds on a PC. Many other classes of strategies have been proposed for MTS ELSPs, but none are analytically tractable for more than a handful of items in the presence of setup times and uncertainties. See Federgruen and Katalan (1996) for a more detailed discussion of the merits of base-stock policies and a review of other classes of policies for MTS systems; see Federgruen and Katalan (1995b) for efficient procedures to search for optimal production schedules.

One of the first models for combined MTS and MTO systems is due to Williams (1984). (See Williams for a review of earlier discussions.) Williams assumes that the MTS items are produced in batches of fixed (but item-dependent) size, requests for which are triggered by (r, Q)-rules. (An (r, Q)-rule triggers a replenishment batch of size Q whenever the item's inventory position drops to a level r.) Orders for the B-items are of random size, and arrive with geometric interarrival times. A dynamic priority rule assigns priority between B- and A-batches (see e.g., Kleinrock 1975). Priority is given to the order or batch with the largest weighted waiting time, where the weight for the

B-items is larger than that applied to all A-items. Among all A-items, as among all B-items, priority is determined on a FIFO basis. Williams approximates the resulting system as one with two classes, both with *Poisson* arrivals. Approximations for the items' lead-time demand distributions and hence optimal reorder levels for given batch sizes Q are obtained from the first (couple of) moments of the waiting times in this queueing system. The end result is an approximate expression for the expected system wide cost as a complex nonlinear function of the batch sizes Q and priority weight u, which is therefore hard to optimize. A similar model for single- and multi-item MTS systems only has been proposed by Karmarkar (1987) and Zipkin (1986), respectively.

The discussion about combined MTS and MTO systems, dormant in the ten years following Williams' paper, was recently revived by Carr et al. (1993) and Sox et al. (1997). To allow for an *exact* and *tractable* analysis, the former authors specialized the Williams' model to one without setup times and setup costs, batches of size one, identical holding and backlogging rates per item, a single production time distribution shared by all A-items, and absolute priority of the B-items over the A-items. To further simplify the analysis, priority is assumed to be preemptive. Under these restrictions, the arrivals of orders to the production facility are indeed generated by a Poisson process, provided the demand processes are themselves independent and Poisson. Sox et al. (1997) consider the same specialized case of the Williams model (in fact, assuming that a *single exponential* production time distribution is shared by *all* items), except that *nonpreemptive* absolute priority is given to the B-items over the A-items. Most recently, Nguyen (1995, 1998) developed heavy traffic limit approximations for various performance measures in hybrid MTO/MTS systems, governed by base-stock policies.

We refer to Federgruen and Katalan (1996) for a review of the literature on pure MTS systems. More recent papers include those by Tayur (1994), which develops analytically exact formulas for the optimal base-stock levels in cyclical base-stock policies, and by Anupindi and Tayur (1998), which proposes a simulation-based method for a variant of cyclical base-stock policies. The few papers within this literature dealing with fully dynamic strategies implicitly encompass hybrid combinations of MTS/MTO strategies. One such example for settings without setup times or costs has been given by Wein (1992). Markowitz et al. (1995) develop heuristics, based on two types of heavy traffic analysis, for systems with setup costs or times.

The remainder of the paper is organized as follows: In §2, we introduce the model and the required notation. In §3, we discuss efficient evaluation methods for the general class of production strategies, described above. In §4, we establish stochastic rankings for several performance measures under various choices for component (i). Section 5 discusses stability conditions for the various priority rules under component (i). Section 6 reports on a numerical study in which the performance of various strategy choices is assessed.

2. The General Model

The production system deals with N distinct items, demands for which are independent and Poisson, with λ_i the demand rate of item i ($i = 1, \ldots, N$). Item 1 is the B-item. Let $\lambda = \sum_{i=1}^{N} \lambda_i$ denote the total demand rate in the system. The results in this paper are easily extended to compound Poisson demand processes. The N items are produced in a common facility that can produce only one unit of one of the items at a time. Production times for individual units are assumed to be independent; those of item i are identically distributed with cdf $S_i(\cdot)$, mean $s_i < \infty$ and kth moment $s_i^{(k)}$ ($k \geq 2$) ($i = 1, \ldots, N$). A possibly random setup time with cdf $R_i(\cdot)$, first moment $r_i < \infty$, and kth moment $r_i^{(k)}$ ($k \geq 2$) is incurred whenever the facility starts producing item i after being idle or after producing some other item. Consecutive setup times are independent. The utilization rate for item i is $\rho_i = \lambda_i s_i$; that of the system equals $\rho = \sum_{i=1}^{N} \rho_i$. We assume the system is stable, i.e., $\rho < 1$. Unfilled demand is backlogged. Three types of costs are incurred. Let

$h_i(x)$ = the inventory carrying cost for item i per unit of time at which x units of item i are carried in stock ($i = 1, \ldots, N$),

$p_i(x)$ = the inventory backlogging cost for item i per unit of time at which x units of item i are backlogged ($i = 1, \ldots, N$), and

K_i = the setup cost incurred per setup of item i ($i = 1, \ldots, N$).

The functions $h_i(\cdot)$ and $p_i(\cdot)$ are convex and nondecreasing. Often, one prefers to control stockouts via service-level constraints rather than explicit backlogging costs, e.g., lower bounds on the items' fill rate. This variant of the basic model calls for a minor adjustment; see §3. We wish to minimize the long run average of total costs.

For any pair of random variables X, Y let $X \oplus Y$ denote the independent sum of X and Y. We write $X = Y$ if the two random variables are identical in distribution. Also, for any nonnegative random variable X with cdf $H(\cdot)$ let $\tilde{H}(\cdot)$ denote the Laplace Stieltjes Transform (LST), $\tilde{H}^{(k)}(\cdot)$ its kth derivative, $k \geq 1$ and $H^e(\cdot)$ the cdf of its forward recurrence time, i.e.,

$$H^e(x) = \int_0^x [1 - H(y)]dy/EX, \quad x \geq 0.$$

For any item-dependent parameter η_i, let $\eta(A) = \sum_{i \in A} \eta_i$.

3. Combined Production Strategies for A- and B-items

Even in the absence of B-items, an optimal production strategy for the A-items (component (ii)) cannot be identified in any reasonable amount of time except for the smallest systems (e.g. with two items) where, in principle, the problem can be solved as a semi-Markov decision problem. Most importantly, the structure of an optimal strategy is highly complex even for special cases of our model, see, e.g., Ha (1992) and Hofri and Ross (1987), precluding its implementation. Instead, we restrict ourselves to the class of base-stock policies which switch between the A-items according to a general periodic sequence. A base stock policy for the A-items is described by:

(a) a vector of base-stock levels $\mathbf{b} = (b_2, \ldots, b_N)$; and

(b) a table $\mathbf{T} = \{T(j);\ j = 1, \ldots, M\}$ of length $M \geq N - 1$, the number of A-items; $T(j)$ denotes one of the items in $A = \{2, \ldots, N\}$ and $T(j) \neq T(j + 1)$(modulo M) for all $j = 1, \ldots, M$.

To describe the policy, assume first that no demands for the B-item arise. At time 0, the facility starts to produce the first item $T(1)$ listed in table $\mathbf{T}$ and continues its production until its inventory level is increased to a base-stock level $b_{T(1)}$. The facility then switches to the second item in the table, $T(2)$, after a setup time $R_{T(2)}$. This protocol continues until the Mth production run for item $T(M)$. Thereafter the facility returns to the beginning of the table, producing its first item $T(1)$ (after a setup time $R_{T(1)}$,) and continuing the above protocol. In particular in the presence of setup costs, it may be beneficial to insert idle times between some of the production runs in $\mathbf{T}$. To simplify the exposition we ignore these idle times at first. In Federgruen and Katalan (1995a) we show how idle times can easily be incorporated.

For a given table $\mathbf{T}$, it is possible to determine an optimal corresponding vector of base-stock levels $\mathbf{b}$, and to evaluate the long-run average costs, via a very efficient solution method described in Federgruen and Katalan (1994). This method starts with the determination of the steady-state shortfall distributions $\{L_1, \ldots, L_N\}$ of the N items, where the shortfall, defined as the difference between the base-stock and the actual inventory level, is independent of the former. Once these shortfall distributions are computed, the problem of determining optimal base-stock levels and of evaluating associated cost measures decomposes into N separate newsboy problems, i.e., N separate minimizations of a convex single-variable function. In the important special case where the holding and backlogging cost functions $h_i(\cdot)$ and $p_i(\cdot)$ are linear, solution of each newsboy problem reduces to that of determining a given fractile of a shortfall distribution.

The method to compute the shortfall distribution uses the fact that each of the shortfalls L_i ($i = 1, \ldots, N$) can be decomposed as the convolution of L_i', the queue size in a dedicated $M/G/1$ system, and an independent component L_i'', i.e., $L_i = L_i' \oplus L_i''$. The distribution of L_i' can be computed exactly, see e.g., Tijms (1986). The method to compute the distribution of L_i'' is exact as well except for approximating the so-called intervisit times by numerically convenient phase type distributions, fitting *any* desired number of moments. The intervisit time I_j for the jth entry in the table denotes the interval of time between the start of the corresponding production run and the termination of the preceding production run for the same item $T(j)$, $j = 1, \ldots, M$. The moments of these intervisit times are themselves computed, within any required relative precision $\epsilon > 0$, via the recursive descendant set method of Konheim et al. (1994). To compute the first m (say) moments of the intervisit times via this method, it suffices to know the first m moments of the production and setup-time distributions. (In practice, it suffices to match two or three moments.) The worst-case complexity of our method for determining optimal base-stock levels and assessing associated cost measures is $O(\max\{Nk^{*2}, M^2 \log_\rho \epsilon\})$ where k^* denotes the largest shortfall level that needs to be evaluated when solving the above newsboy problems. The complete method requires only a few seconds on a PC and is remarkably accurate as verified in an extensive numerical study. See Appendix 1 in Federgruen and Katalan (1995a) for a more detailed summary of this procedure. Finally, the above method can be applied to any given table **T** or any menu of such tables. See Federgruen and Katalan (1995b) for a systematic optimization over all possible tables **T**.

This base-stock policy needs to be complemented with an *interruption* discipline (component (i)) to allow for quick response production of the B-item. In its purest form, no inventory is kept for the B-item. More generally we may maintain a small base-stock level b_1. Several interruption disciplines may be envisioned. A first important distinction is how *production* of the A-items is interrupted:

(I) Absolute Priority rules: Absolute priority is given to orders for B-units: either (PS), *preemptive-resume*, immediately when the order is placed, or (NP), *nonpreemptive*, as soon as the unit currently being produced is completed. Under (PS) we assume that no work is lost due to preemptions. See §10 in Federgruen and Katalan (1995a) for a treatment of variations where some or all of the work on a unit in process is lost.

(II) Postponable Priority rules: Here, production of B-units is inserted into the production schedule of the A-items, but only when the facility would otherwise switch between A-items. We refer to these as (PP)-rules.

Absolute priority rules provide quicker response to the B-item than postponable priority rules but appear attractive only when the desired service level provided to the B-item or the backlogging cost associated with this item is significantly larger than those of the A-items or when the time and cost required to setup for an A-item after having worked on the B-item is negligible or relatively small. To simplify the exposition, we assume that the same characteristic (PS, NP, PP) applies to all A-items; extensions to mixed settings are straightforward.

Assuming that B-units are given absolute priority, it is clearly advantageous to preempt the setup for an A-item as soon as an order for a B-unit arrives. As with the production times, we assume that no work is lost due to preemption of setups. For alternative assumptions, see §10 in Federgruen and Katalan (1995a).

Under a *postponable* rule, the B-item is inserted once or multiple times as an entry into the **T**-table, giving rise to a new extended **T**-table for which an associated optimal base-stock policy can be determined via the general method described above. We now show that under *absolute* priority rules, performance of the A- and B-items can be ascertained respectively from (i) that of an equivalent system *without* the B-item, but with modified setup and production time distributions; and (ii) a simple "$M/G/1$ system with vacations," see, e.g., Doshi (1990).

We may consider certain dynamic adjustments to the above class of policies, e.g. where an item is skipped when it is its turn to be produced but its inventory is still at its base-stock level. This adjustment would apply to any of the A-items under absolute priority rules and to all items, (the B-item included), under postponable rules. Günalay and Gupta (1997) provide two numerical methods to accommodate this dynamic adjustment. Their first method is an adaptation of the above-mentioned descendent set method by Konheim et al. (1994). Unfortunately, it applies only to systems with two items. The second method, based on Discrete Fourier Transforms, applies to an arbitrary number of items but it is computationally intractable except when the number of items is small and $\rho < 0.5$. Also, the event under which an item can be skipped, when following the table **T**, because its inventory has not dropped from its base-stock level, is rare under reasonably large utilization rates and/or cycle times. This implies that the performance measures obtained without the dynamic adjustment may be used as a good approximation.

Absolute Priority Rules

For $i = 2, \ldots, N$, let $\hat{R}_i$ denote the total amount of time between the beginning of a *setup* for item i and its completion, and $\hat{S}_i$ the total time between the beginning of the *production* of a unit of item i and its conclusion, including possible interruptions to produce the B-item. Note that as far as the A-items are concerned, the system is equivalent to one without B-units, and $\{\hat{R}_i, \hat{S}_i; i = 2, \ldots, N\}$ as the setup times and unit production times. Recall that the evaluation and optimization method described above requires knowledge of a given number of moments of *all* setup and production time distributions. Highly accurate approximations can be obtained on the basis of the first two or three moments only. We now derive the marginal distributions $\{\hat{R}_i, \hat{S}_i; i = 2, \ldots, N\}$. We first introduce some notation. Let

$$B = \text{the length of a busy period for the B-item in a dedicated system for this item,} \quad (1)$$

i.e., a *dedicated* $M/G/1$ system with exceptional first service, which has arrival rate λ_1, regular service time distribution $S_1(\cdot)$, and first service time distribution $R_1 \oplus S_1$. The moments of B are obtained from the well-known identity relating its LST $\tilde{B}(\cdot)$ to $\tilde{S}_1(\cdot)$ and $\tilde{R}_1(\cdot)$ (cf., e.g., Wolff 1989):

$$\tilde{B}(s) = \tilde{R}_1[s + \lambda_1(1 - \tilde{B}_0(s))]\tilde{B}_0(s) \quad \text{with } \tilde{B}_0(s) = \tilde{S}_1[s + \lambda_1(1 - \tilde{B}_0(s))]. \quad (2)$$

Closed form expressions for the first three moments are given in Appendix 2 of Federgruen and Katalan (1995a). Below, $\{B_1, B_2, \cdots\}$ refers to a sequence of i.i.d. random variables, distributed like B. Also, for any interval of time U, let

$$N(U) = \text{the number of B-units demanded during } U \text{ or an interval of time distributed as } U. \quad (3)$$

Clearly,

$$\hat{R}_i = R_i + \sum_{l=1}^{N(R_i)} B_l. \quad (4)$$

To obtain a qualitative understanding of the impact B-units have on the "extended" setup times for the A-items, consider the case where $R_1 = 0$. Proposition 1 in Federgruen and Katalan (1995a) shows that

$$\mathrm{Var}(\hat{R}_i) = (\lambda_1 r_i s_1^{(2)} + \mathrm{Var}(R_i)(1 - \rho_1))/(1 - \rho_1)^3,$$

so that the squared coefficient of variation

$$\mathrm{CV}^2(\hat{R}_i) = \mathrm{CV}^2(R_i) + \frac{\rho_1}{1 - \rho_1} \frac{s_1(1 + \mathrm{CV}^2(S_1))}{r_i}.$$

The second term in this expression thus denotes the *increase* in the squared coefficient of variation due to interruption for B-units. In other words, setup times always become more variable due to interruptions for the MTO items. The increase grows to infinity with ρ_1, the fraction of the capacity required to produce MTO items, and when production times of the B-units are relatively large or variable compared to the mean interrupted setup time. Interestingly, the increase in variability is particularly large when mean setup times without interruptions by B-orders are *small*, i.e., $\mathrm{CV}^2(\hat{R}_i) \to \infty$ as $r_i \downarrow 0$. This arises because under interruptions for the B-item, the extended setup times are considerably larger than in the event that no demand for the B-item arises during the initial setup time. These observations indicate that providing absolute priority to a large fraction of the orders can result in significant increases of the shortfall distributions for the A-items, or the inventories required to support given service levels beyond what can be expected from the mere increase in their *expected* cycle times $\{C_2, \ldots, C_N\}$, i.e., the time between consecutive production runs for a given item.

The modified *production* time distributions in the equivalent system depend on which of the characteristics (PS) or (NP) prevails.

$$\text{(PS):}\quad \hat{S}_i = S_i + \sum_{l=1}^{N(S_i)} (B_l + \hat{R}_{i,l})$$

$$\text{where } \hat{R}_{i,1}, \hat{R}_{i,2}, \cdots \text{ are i.i.d. as } \hat{R}_i, \quad (5)$$

$$\text{(NP):}\quad \hat{S}_i = S_i + \sum_{l=1}^{N(S_i)} B_l + \begin{cases} \hat{R}_i & \text{if } N(S_i) > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

The shortfall distributions of the A-items and associated base-stock levels can be computed by applying the general method described above to the equivalent system with its modified setup and production time distributions, cf., (4)–(6). As far as item 1 is concerned, its cost under a given (e.g., *zero*) base-stock level is again obtained from the item's steady-state shortfall distribution L_1. We write L_1(PS), L_1(NP), and L_1(PP) to distinguish between the shortfall distribution under preemptive priority (PS), nonpreemptive priority (NP), and postponable priority (PP), respectively. Similarly, we write W_1(PS), W_1(NP), and W_1(PP) for the steady-state waiting time for a B-order under the three types of priority rules.

Under *preemptive* priority, i.e., (PS), the shortfall L_1 is distributed as L'_1, the steady-state queue size in the dedicated $M/G/1$ system with exceptional first service time. Under the two alternative types of priority, L_1 is the queue size in this $M/G/1$ system, interrupted by "generalized vacations." The system is "on vacation" whenever the facility is producing or setting up for A-items. Note that a vacation starts as soon as item 1's shortfall is reduced to zero. Under (NP) priority, it terminates upon the first demand for a B-unit, if the facility is setting up at that time and at the first subsequent production completion epoch, otherwise. Under (PP) priority, the vacation terminates as soon as the facility switches back from producing an A-item to that of the B-item. In both cases, the vacation structure satisfies assumptions (1)–(6) in Fuhrmann and Cooper (1985). By their decomposition result, we have that L_1(NP) and L_1(PP) may be written as a convolution of L'_1 and an independent component L''_1(NP) and L''_1(PP), respectively. Similarly, W_1(NP) and W_1(PP) may be written as a convolution of W'_1, the steady-state waiting time in the dedicated $M/G/1$ system without interruptions, and an independent component W''_1(NP) and W''_1(PP), respectively. While L''_1(NP) and W''_1(NP) can be characterized as simple mixtures of elementary random variables, such a characterization is more complex in the case of L''_1(PP) and W''_1(PP). See the discussion in the introductory part of this section, and, for more details, see Federgruen and Katalan (1994). We thus obtain:

THEOREM 1.
(*a*) (*i*) $L_1(PS) = L'_1$;
(*ii*) $L_1(NP) = L'_1 \oplus L''_1(NP)$ *where* $L''_1(NP)$ *is a mixture of* $\{N(S^e_2), N(S^e_3), \ldots, N(S^e_N), 0\}$ *with mixing probabilities* $\rho_2/(1 - \rho_1), \rho_3/(1 - \rho_1), \ldots, \rho_N/(1 - \rho_1), (1 - \rho)/(1 - \rho_1)$;
(*iii*) $L_1(PP)$ *can be decomposed as the convolution of* L'_1 *and an independent nonnegative random variable* $L''_1(PP)$, *i.e.,* $L_1(PP) = L'_1 \oplus L''_1(PP)$.
(*b*) (*i*) $W_1(PS) = W'_1$;
(*ii*) $W_1(NP) = W'_1 \oplus W''_1(NP)$ *where* $W''_1(NP)$ *is a mixture of* $\{S^e_2, S^e_3, \ldots, S^e_N, Z\}$ *with* Z *an exponentially distributed random variable with mean* λ_1^{-1}, *and mixing probabilities* $\rho_2/(1 - \rho_1), \rho_3/(1 - \rho_1), \ldots, \rho_N/(1 - \rho_1), (1 - \rho)/(1 - \rho_1)$;
(*iii*) $W_1(PP)$ *can be decomposed as the convolution of* W'_1 *and an independent nonnegative random variable* $W''_1(PP)$, *i.e.,* $W_1(PP) = W'_1 \oplus W''_1(PP)$.

PROOF. Except for the characteristics above of L''_1(NP) and W''_1(NP), the theorem follows from Fuhrmann and Cooper (1985) and the discussion above. We now verify the characterization of L''_1(NP), the proof of that W''_1(NP) being analogous.

It follows from Fuhrmann and Cooper (1985) that L''_1(NP) is the queue size at an arbitrary tagged epoch during one of the vacations. Let E_i denote the event that the tagged epoch falls within a production period for item *i*, $i = 2, \ldots, N$. Recall that under E_i and by the definition of rule (NP), the queue is empty at the onset of the unit production period in which the tagged epoch falls and the steady-state distribution of the age of this unit production time is distributed as S^e_i, see, e.g., Wolff (1988). It follows that $(L''_1(\text{NP}) \mid E_i) \overset{d}{=} N(S^e_i)$. Also, we have $(L''_1 \mid$ the tagged epoch occurs when the system is setting up$) = 0$, and the probabilities of the conditioning events $\{E_i: i = 2, \ldots, N\}$ satisfy $\Pr[E_i] = \Pr[\text{system is producing item } i \mid \text{system is not producing item } 1] = \Pr[\text{system is producing item } i]/\Pr[\text{system is not producing item } 1] = \rho_i/(1 - \rho_1)$. □

4. Comparison Between Alternative Priority Rules

In this section we compare the performance of the A- and B-items under absolute and postponable priority rules. Theorem 2 below compares the shortfall and order delay distributions for the B-item assuming for simplicity's sake that this item is purely MTO, i.e., no inventory is kept for it.

THEOREM 2.
(*a*) $L_1(PS) \le_{st} L_1(NP) \le_{st} L_1(PP)$,
(*b*) $W_1(PS) \le_{st} W_1(NP) \le_{st} W_1(PP)$.

PROOF. (a) The first inequality is immediate from Theorem 1, part (a). $L''_1(\text{PP})$ represents the shortfall for item 1 at an arbitrary tagged epoch during an intervisit time for this item. As in the proof of Theorem 1, let E_i denote the event that this tagged epoch falls within a production run of type i, $i = 2, \ldots, N$ and let E_0 denote the complementary event. As shown in the proof of Theorem 1, $\Pr[E_i] = \rho_i/(1 - \rho_1)$ for $i = 2, \ldots, N$ and $\Pr[E_0] = (1 - \rho)/(1 - \rho_1)$. Note that $(L''_1(\text{PP}) \mid E_0) \ge 0$ a.s. For any $i = 2, \ldots, N$ $(L''_1(\text{PP}) \mid E_i)$ is almost surely larger than the number of units of item 1 demanded during the elapsed part of the unit production time in process at the tagged epoch, which is distributed as S_i^e, cf. Wolff (1988). Thus, $(L''_1(\text{PP}) \mid E_i) \ge_{st} N(S_i^e) = (L''_1(\text{NP}) \mid E_i)$; see the proof of Theorem 1. Hence $L''_1(\text{PP}) \ge_{st} L''_1(\text{NP})$, since the mixing probabilities are equal under both rules, and $L_1(\text{PP}) = L'_1 \oplus L''_1(\text{PP}) \ge_{st} L'_1 \oplus L''_1(\text{NP}) = L_1(\text{NP})$.

(b) It follows from Fuhrmann and Cooper (1985) (see also Federgruen and Katalan (1994, pp. 362–363)) that $W''_1(\text{PP})$ and $W''_1(\text{NP})$ are distributed as the age of the intervisit time in process at an arbitrary tagged epoch during an intervisit time. The remainder of the proof is analogous to that of part (a). □

We conclude that the rules (PS), (NP), and (PP) generate progressively larger shortfalls and order delays for the B-item. Moreover, these can be characterized as a common basic shortfall or order delay, experienced in a system fully dedicated to these items, plus a progressively larger component that thus can be viewed as the penalty paid for the restricted priority attributed to the B-item. For example, Theorem 1 shows that the increase in the mean shortfall under *nonpreemptive* absolute priority (NP) is given by

$$\sum_{i=2}^{N} \frac{\rho_i}{(1-\rho_1)} \frac{\lambda_1 s_i^{(2)}}{2s_i} = \frac{\lambda_1^2 s_1^{(2)}}{2(1-\rho_1)} \sum_{i=2}^{N} \frac{\lambda_i s_i^{(2)}}{\lambda_1 s_1^{(2)}}. \tag{7}$$

The first factor to the right of (7) represents, under the preemptive rule (PS), the expected number of B-units waiting to be processed, i.e., the expected queue size in the uninterrupted $M/G/1$ system. The second factor is a dimensionless index; e.g. when (the second moments of) all unit production times are identical, the index equals the ratio of the aggregate demand rate of the A-items and that of the B-item. Thus a *relatively* high price is paid in terms of the cost incurred for the B-item when switching from preemptive to nonpreemptive priority, in particular when demand for the B-item is relatively small.

One would expect that the progressively larger shortfalls and order delays for the B-item under the above sequence of priority rules is compensated by progressively shorter intervisit and cycle times experienced by the A-items. Proposition 1 below establishes this stochastic ranking of the absolute priority rules for *cyclical* base-stock production strategies. For all $i = 2, \ldots, N$, let $I_i(\text{PS})$ and $I_i(\text{NP})$ denote the intervisit time of item i when applying the PS- and NP-rule, respectively. Similarly, for all $i = 2, \ldots, N$, let $C_i(\text{PS})$ and $C_i(\text{NP})$ denote the cycle time when applying the PS- and NP-rule, respectively.

PROPOSITION 1. *For all* $i = 2, \ldots, N$, $I_i(PS) \ge_{st} I_i(NP)$; $C_i(PS) \ge_{st} C_i(NP)$.

PROOF. Without loss of generality let $i^* = 2$. Altman et al. (1992, Prop. 4.3) show for cyclical polling systems that if one of the service time distributions is replaced by a stochastically larger one, *all* intervisit and cycle times of all items are stochastically increased as well. It thus suffices to verify that the extended production times are appropriately stochastically ranked. The result is immediate since $\sum_{l=1}^{N(S_i)} \hat{R}_{i,l} \geq 1\{N(S_i) > 0\}\hat{R}_i$ a.s. □

REMARK. Let I_i(PP) and C_i(PP) denote the intervisit time and cycle time for item i under the postponable priority rule in which item 1 is inserted after every one of the entries of the cyclical table for the A-items. One may surmise that the range of stochastic comparisons could be extended to:

$$I_i(\text{PS}) \geq_{st} I_i(\text{NP}) \geq_{st} I_i(\text{PP}) \quad \text{and}$$
$$C_i(\text{PS}) \geq_{st} C_i(\text{NP}) \geq_{st} C_i(\text{PP})$$
$$\text{for all } i = 2, \ldots, N.$$

This, however, fails to be true. For example, under a sufficiently low demand rate for item 1, the PP-rule may insert more setup times for this item in any given cycle than the absolute priority rules. (Recall, the PP-rule prescribes a setup for item 1 after the completion of each production run for any of the A-items regardless of whether the inventory of the B-item is at its base-stock level or not.) It remains an open question whether the stochastic ordering applies to $\overline{\text{PP}}$, the dynamic adjustment of PP, under which a setup of the B-item is skipped if its prevailing inventory level is still at its base-stock level. Switching from the nonpreemptive absolute priority rule (NP) to the $\overline{\text{PP}}$-rule (stochastically) decreases the intervisit and cycle times of the first cycle but it remains an open question whether the same ranking carries over to the equilibrium distributions.

5. Stability Conditions

The choices for component (i), in particular (PS), (NP), and (PP) have implications for the stability of the system. As mentioned in §2, the inequality $\rho = \sum_{i=1}^{N} \rho_i < 1$ is clearly a *necessary* condition for stability. Under alternative (PP), it is sufficient as well, see Takagi (1986, 1990). On the other hand, under absolute priority rules, the system may be unstable even if $\rho < 1$ and even if B-items can be inserted without setup times. In fact the stability condition in this case is given by

$$\sum_{i \in A} \lambda_i E\hat{S}_i < 1, \tag{8}$$

with $\hat{S}_i$, the extended unit production time, specified as in (5) and (6).

The proof of Proposition 1 shows that for all $i \in A$ $\hat{S}_i(\text{PS}) \geq_{st} \hat{S}_i(\text{NP})$ and in particular $E(\hat{S}_i(\text{PS})) \geq E(\hat{S}_i(\text{NP}))$. Thus, stability may in particular be jeopardized when adopting preemptive priority rules for all items. In this case the stability condition (8) can be stated as follows:

PROPOSITION 2. *Under* (*PS*) *the system is stable if and only if*

$$\sum_{i=2}^{N} \rho_i[1 + \lambda_1 r_1][1 + \lambda_1 r_i]/(1 - \rho_1) < 1$$

or equivalently,

$$\rho < 1 - \sum_{i=2}^{N} \rho_i[\lambda_1 r_i + \lambda_1 r_1(1 + \lambda_1 r_i)]. \tag{9}$$

PROOF. It suffices to verify that (9) is equivalent to (8). To find an expression for $E\hat{S}_i$, note first that $E(B) = (r_1 + s_1)/(1 - \rho_1)$, the well known formula for the expected busy period in an $M/G/1$ system with exceptional first service time; see also (2). It then follows from (4) that

$$E(\hat{R}_i) = r_i + \lambda_1 r_i E(B) = r_i(1 + \lambda_1 r_1)/(1 - \rho_1),$$

and hence from (5)

$$E(\hat{S}_i) = s_i + \lambda_1 s_i[(r_1 + s_1) + r_i(1 + \lambda_1 r_1)]/(1 - \rho_1). \tag{10}$$

Substituting (10) into (8), we obtain Equation (9). □

The second term to the right of Equation (9) represents the fraction of the capacity which is wasted due to setups associated with interruptions of the production of A-units:

$$\text{wasted capacity} \stackrel{\text{def}}{=} \sum_{i=2}^{N} \rho_i[\lambda_1 r_i + \lambda_1 r_1(1 + \lambda_1 r_i)].$$

The wasted capacity is particularly large when the demand rate for the B-item (λ_1) is large, independent of the amount of work associated with the production of these B-units, i.e., independent of S_1. The wasted capacity *does* increase proportionally with the mean production times of the A-items and their demand rates. The stability condition (9) implies for given load factors $\rho_2, \ldots, \rho_N$, that the maximum (sales) volume assignable to the MTO item is distinctly smaller under absolute priority rules (PS) as compared to postponable ones (PP). Such limitations cannot be observed in models without setup times, in which it is sometimes optimal to assign close to 100% of the load to the B-items, see Carr et al. (1993).

Also, the utilization rate under preemptive interruptions for the B-item equals

$$\sum_{i=2}^{N} \rho_i[1 + \lambda_1 r_1][1 + \lambda_1 r_i]/(1 - \rho_1),$$

see above. Giving preemptive priority to the B-item always increases the mean cycle times experienced by the A-items, since in the modified system

$$E\hat{C} = \sum_{j=1}^{M} E(\hat{R}_{T(j)}) \Bigg/ \left(1 - \sum_{i=2}^{N} (\lambda_i E\hat{S}_i)\right)$$

$$\geq (1 - \rho_1)^{-1}\left(\sum_{j=1}^{M} r_{T(j)}\right) \Bigg/ \left[1 - (1 - \rho_1)^{-1} \sum_{i=2}^{N} \rho_i\right]$$

$$= \sum_{j=1}^{M} r_{T(j)}/(1 - \rho) = EC$$

where the inequality follows from $E\hat{S}_i \geq s_i(1 + \lambda_i EB) = s_i/(1 - \rho_1)$ by Proposition 3 (Federgruen and Katalan 1995a, Appendix 2).

6. Numerical Study

We have conducted a numerical study to investigate some of the main choices presented in this paper. In particular, we have gauged whether *absolute* priority rules are to be preferred over postponable ones, and if so under what circumstances and via what specific absolute or postponable priority scheme. We also apply our methods to characterize the logistical implications of adding a low volume specialized B-item to a given product line. These implications include the allocation mechanism of the production capacity among MTO and MTS items and the increase in inventory needs for the basic A-items.

To evaluate the relative performance of various absolute and postponable priority rules, we focus on settings where a class of B-units and its service with absolute priority rules is most likely to be effective, i.e., where the B-items have zero setup times ($r_1 = 0$) and *all* items have zero setup costs. We show that even in this extreme case, postponable priority rules often dominate absolute priority rules. When $r_1 > 0$, it can be expected that postponable rules start to dominate for even lower setup times for the A-items, or for an even lower demand volume or penalty cost rate for the B-item, etc. Similarly, one can expect that the optimal frequency of inserting the B-item within the table **T**, (among all considered postponable rules), decreases as the value of r_1 increases. These monotonicities are based on the fact that the incremental wasted capacity under absolute priority rules beyond that incurred under the postponable rules, increases as item B's setup time r_1 increases. See also the discussion in §5.

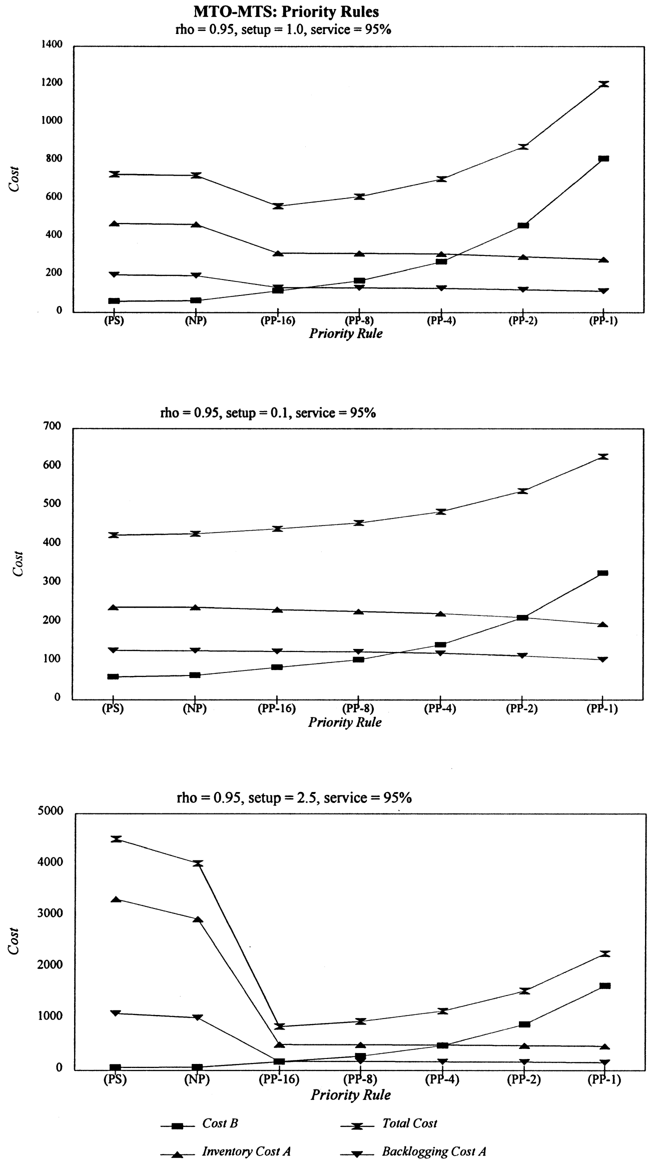
We start by investigating a class of 675 problem instances, all with 16 A-items, where the demand volume of the A-class as a percentage of the overall demand varies between 70% to 90%. This type of breakdown of the product line is consistent with classical A/B/C classifications. A power-of-two number of A-items is chosen to easily investigate the impact of doubling the frequency of service to the B-item via (PP) rules.

Assume, without loss of generality, that the A-items are ordered in increasing order of their demand values. Consistent with empirically observed Pareto-curves, we have constructed these to be of exponential shape for the A-class, i.e., $\lambda_{i-1}/\lambda_i = q$ for some constant $q < 1$, for all $i = 3, \ldots, 17$ A-items. All items have the same unit production time (exponential with mean 1), and holding and penalty cost rates ($h_i = h = 1$ and $p_i = p$ for all $i = 1, \ldots, 17$). We assume however that the B-item cannot be stocked. All A-items have identical exponential setup time distributions. In particular, $r_i \stackrel{\text{def}}{=} r$ for all $i = 2, \ldots, 17$. (Recall that the B-item has zero setup time.) The 675 problem instances are obtained by combining (i) 5 possible choices of $\rho = \sum_{i=1}^{17} \lambda_i$, i.e., $\rho = 0.6, 0.7, 0.85, 0.9, 0.95$, (ii) 3 possible values for the *shape parameter* q of the Pareto-curve, i.e., $q = 0.9, 0.95, 0.99$, (iii) 3 possible values for $\rho(A)/\rho = \lambda(A)/\sum_{i=1}^{17} \lambda_i = 0.7$, 0.8 and 0.9, (iv) 3 possible values for $r = 0.1, 1, 2.5$, and (v) 5 different values for the ratio $p/(p + h) = 0.25, 0.5, 0.9, 0.95, 0.99$. All demand rates are fully specified by the triple $(\rho, \rho(A), q)$.

We assume that the A-items are governed by a *cyclical* base-stock policy, i.e., $\mathbf{T} = \{2, \ldots, 17\}$. For all 675 problem instances, we have evaluated both the preemptive and nonpreemptive priority rules (PS) and (NP) for all 16 A-items as well as 5 distinct postponable rules (PP-2^i, $i = 0, \ldots, 4$), i.e., where the B-class is inserted into the table $\mathbf{T}$ 2^i times after every 2^{4-i} A-items. Both the restriction to cyclical policies for the A-items and the restriction to the five possibilities (PP-2^i) are made for the sake of brevity only. As before, we refer the reader to Federgruen and Katalan (1995b) for a systematic discussion of search methods for an optimal (extended) table, $\mathbf{T}$.

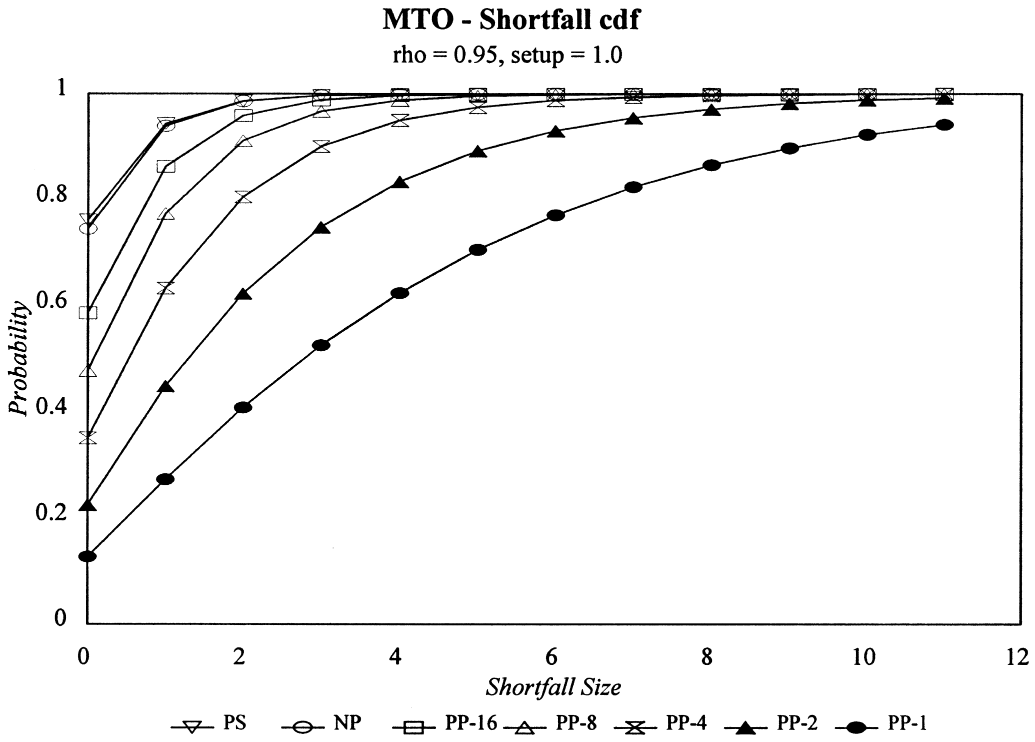
The "stability problem" discussed in §5 arises quite frequently, even though zero setup times are assumed for the B-item. Recall that the wasted capacity measure in (9) is minimized when $r_1 = 0$. As can be expected from the wasted capacity term in (9), the stability problem arises in particular for instances with a high utilization rate ρ and large setup times. Even when the system continues to be stable under absolute priority rules, we have found that these are often dominated in terms of overall cost performance by one of the five postponable rules, and as demonstrated below, performance differences can be highly significant. Which of the seven considered priority rules dominates, and to what extent, depends on the specific parameters; absolute priority rules tend to dominate when setup times are low, the demand rate for the B-class is small, and a high service level $(p/(p + h))$ is required. Even when an absolute priority rule is used, it may be of significant advantage to support the B-item with its own base-stock in settings, different from the above scenarios, where the B-item can be stocked. This applies in particular when a high service level needs to be guaranteed, and when the demand rate for the B-item is not too small. More subtle changes in the relative ranking of the priority rules can be expected under non-identical parameters and distributions. We conclude that a *systematic* optimization is required for the different choices for each of the strategy components rather than a categorical, a priori restriction to *one* of them.

Figure 1 **MTO—MTS: Priority Rules**



We illustrate the above with the help of a few graphs for a few specific problem instances from the

Figure 2 **MTO—Shortfall cdf**
MTS—MTO: Base Stock Level—A Class



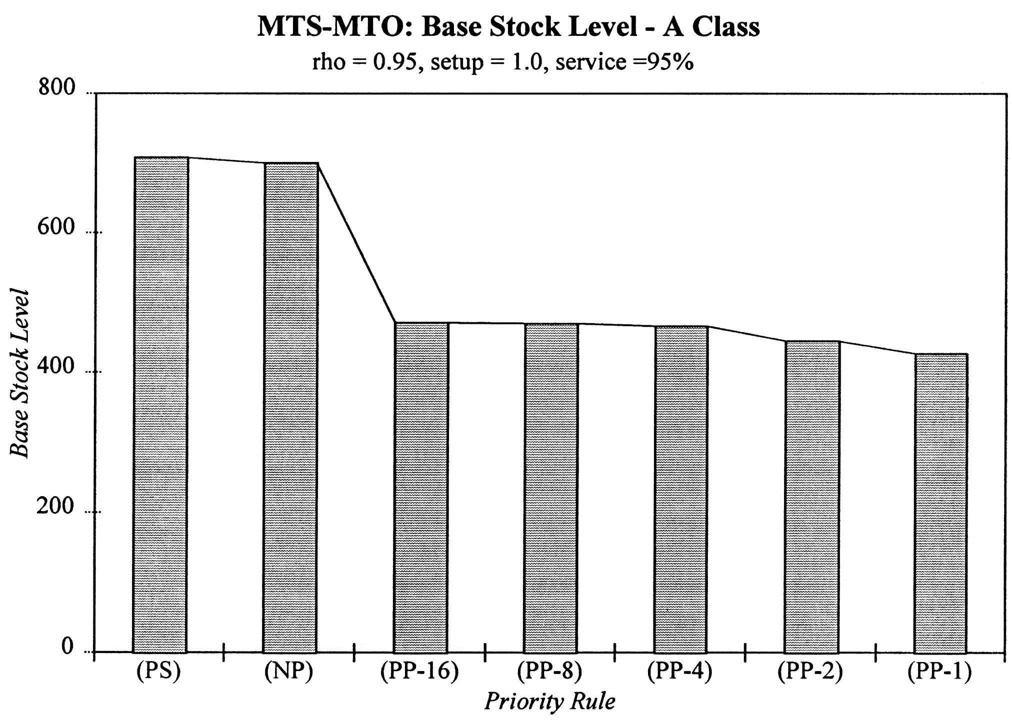
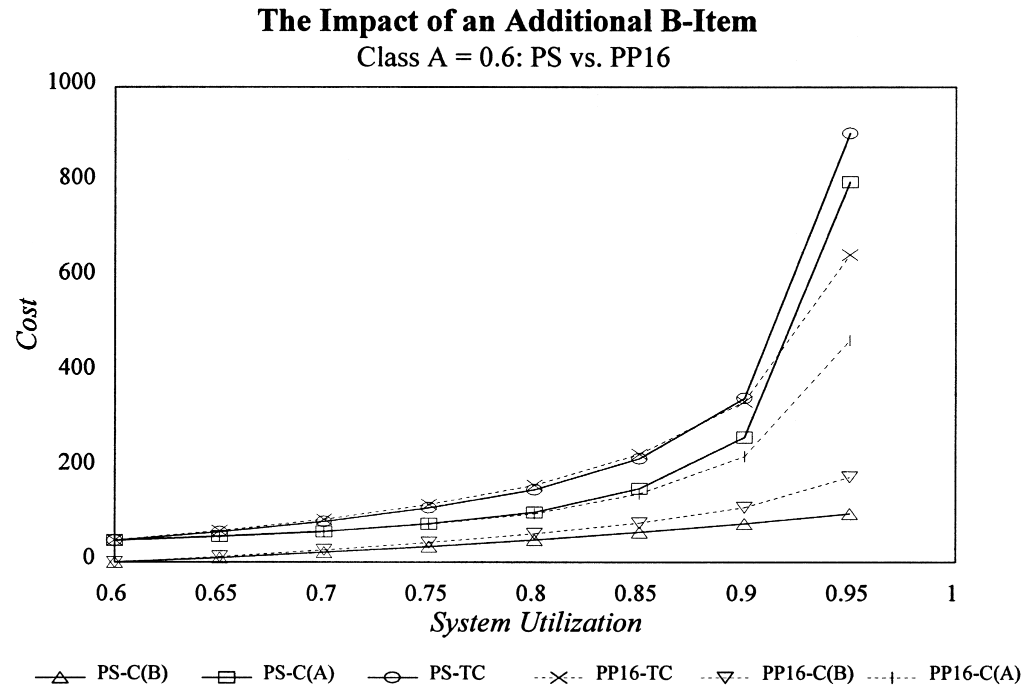
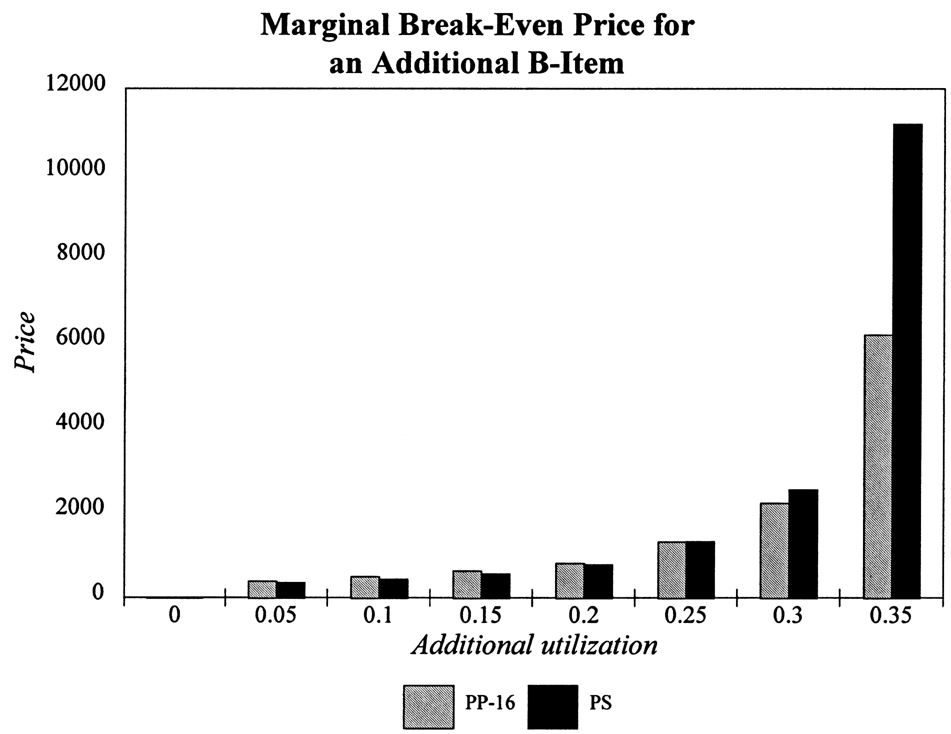


Figure 3 The Impact of an Additional B-Item
Marginal Break-Even Price for an Additional B-Item





above class of 675 in which the mean unit production time for the B-item as well as its p- and h-values are increased by a factor of ten and its demand rate decreased by the same factor, leaving the utilization rate unaltered. This modification was made to create instances in which all of the considered priority rules are feasible, i.e., result in stable systems; see our discussion above. The selected problem instances all have $q = 0.95$, $p/(p + h) = 0.95$ and $\rho(A)/\rho = 0.75$. The *basic* instance considered has $r = 1$ and $\rho = 0.95$. (This setup time value can in many settings be viewed as moderate since it equals the mean time required to produce a *single* unit.)

Figure 1 compares the performance of the seven priority rules, ranked in decreasing order of priority provided to the B-class, for the basic instance and two others in which the mean setup times are changed from 1 to 0.1 and 2.5, the three values considered in the full set of 675 instances. (Even a mean setup time of 2.5 is moderate since the average production run of the A-items in this instance takes approximately 35 time units under any of the postponable rules.) In addition to the total cost, we depict its three principal components, the cost for the B-items, the holding cost for the A-items and their backlogging costs. We assume that the B-item is truly MTO, i.e., no inventory can be maintained for this item. We observe that for the basic instance, (PP-16) is optimal, significantly improving upon, e.g., either one of the two extremes (PS) or (PP-1).

When the mean setup time is reduced to 0.1, preemptive priority (PS) becomes optimal; when it is increased to 2.5, (PP-16) continues to be optimal. In this case, (PP-16) is *five* times more efficient than the absolute priority rule (PS), while it continues to be twice as efficient as the other extreme (PP-1).

Figure 2 displays, for the basic instance, the cdfs of L_1, the aggregate shortfall distribution of the B-items, under the seven rules; these exhibit the stochastic rankings stated in Theorem 2. The bottom part of this figure exhibits the optimal total base-stocks under the seven rules; these represent a decreasing curve, consistent with the stochastically decreasing cycle times, proved in Proposition 1, for the absolute priority rules.

Figure 3 addresses the strategic question mentioned above. We modify the basic instance, in which $\rho(A)$ is fixed at $\rho(A) = 0.6$ and ρ is progressively increased from $\rho = 0.6$ to $\rho = 0.95$ by adding a low-volume specialized B-item to the basic product line. The upper part of the figure exhibits for the absolute priority rule (PS) and the postponable rule (PP-16), how the total expected costs as well as the cost for the A-items and the B-item separately, depend on the amount of additional business of B-units. The absolute (postponable) priority rule dominates for relatively low (high) utilization rate ρ, i.e., $\rho < 0.9$ ($\rho \geq 0.9$). While the cost increases for the B-item remain moderate, those incurred for the A-items and hence for the total system costs, become severe when ρ increases beyond 85%, say.

These cost curves allow us to calculate a marginal break-even price at each of the considered additional utilization rates by computing the total cost differentials associated with increments of ρ by 5 percentage points at a time. The marginal break-even prices are displayed at the bottom part of Figure 3 and can be used as input to sales negotiations. Note that the marginal break-even price is roughly 100 times larger at $\rho = 0.9$ than at $\rho = 0.6$, indicating how much caution is to be exercised when accepting additional specialized business at a high utilization rate, and how misleading standard cost accounting methods can be.

References

Altman E., P. Konstantopoulos, Z. Liu. 1992. Stability, monotonicity and invariant quantities in general polling systems. *Queueing Systems* **11** 35–57.

Anupindi R., S. Tayur. 1998. Managing stochastic multi-product systems: Models, measures and analysis. *Oper. Res.* **46** S98–S111.

Carr A. S., A. R. Güllü, P. L. Jackson, J. A. Muckstadt. 1993. An exact analysis of a production-inventory strategy for industrial suppliers. Working Paper. School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.

Doshi B. T. 1990. Single server queues with vacations. H. Takagi (ed.), *Stochastic Analysis of Computer and Communication Systems*, North-Holland, Amsterdam, 217–265.

Federgruen A., Z. Katalan. 1994. Approximating queue size and waiting time distributions in general polling systems. *Queueing Systems* **18** 353–386.

——, ——. 1995a. Make-to-stock or make to order: that is the question; novel answers to an ancient debate. Unabridged version of this paper, Graduate School of Business, Columbia University, New York.

——, ——. 1995b. Determining production schedules under base-stock policies in single facility multi-item production systems. Forthcoming in *Oper. Res.*

——, ——. 1996. The stochastic economic lot scheduling problem: cyclical base-stock policies with idle times. *Management Sci.* **42** 783–796.

Fuhrmann S. W., R. B. Cooper. 1985. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Oper. Res.* **33** (5) 1117–1129.

Günalay Y., D. Gupta. 1997. Polling systems with a patient server and state-dependent setup times. *IIE Trans.* **29** 469–480.

Ha A. 1992. Optimal scheduling and stock rationing policies in make-to-stock production systems. Ph.D. dissertation, Stanford University, Stanford, CA.

Hofri M., K. W. Ross. 1987. On the optimal control of two queues with server setup times and its analysis. *SIAM J. Comput.* **16** 399–420.

Karmarkar U. S. 1987. Lot sizes, lead times and in-process inventories. *Management Sci.* **33** 409–418.

Kleinrock L. 1975. *Queueing Systems*, vol. 1. John Wiley & Sons, New York.

Konheim A. G., H. Levy, M. M. Srinivasan. 1994. Descendant set: An efficient approach for the analysis of polling systems. *IEEE Trans. Comm.* **42** (2/3/4) 1245–1253.

Markowitz D. M., M. I. Reiman, L. M. Wein. 1995. The stochastic economic lot scheduling problem: heavy traffic analysis of dynamic cyclic policies. Working Paper. Operations Research Center, MIT, Cambridge, MA.

Nguyen V. 1995. On base-stock policies for make-to-order/make-to-stock production. Working Paper. Sloan School of Management, MIT, Cambridge, MA.

——. 1998. A multiclass hybrid production center in heavy traffic. *Oper. Res.* **46** (3) S13–S25.

Perkins A. G. 1994. Manufacturing. *Harvard Bus. Rev.* **72** (2) 13–14.

Sox C. R., L. J. Thomas, J. O. McClain. 1997. Coordinating production and inventory to improve service. *Management Sci.* **43** (9) 1189–1197.

Takagi H. 1986. *Analysis of polling systems*. MIT Press, Cambridge, MA.

——. 1990. Queueing analysis of polling models: an update. H. Takagi (ed.), *Stochastic Analysis of Computer and Communication Systems*, North-Holland, Amsterdam, 267–318.

Tayur S. 1994. A simple formula is asymptotically exact for certain complex problems. Working Paper. GSIA, Carnegie Mellon, Pittsburgh, PA.

Tijms H. C. 1986. *Stochastic Modeling and Analysis: A Computational Approach*. John Wiley & Sons, UK.

Wein L. 1992. Dynamic scheduling of a multi-class make-to-stock queue. *Oper. Res.* **40** 724–735.

Williams T. M. 1984. Special products and uncertainty in production/inventory systems. *European J. Oper. Res.* **15** 46–54.

Wolff R. W. 1988. Sample-path derivations of the excess, age, and spread distributions. *J. Appl. Probab.* **25** 432–436.

——. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, NJ.

Zipkin P. 1986. Models for design and control of stochastic, multi-item batch production systems. *Oper. Res.* **34** (1) 91–104.

Accepted by Hau Lee; received October 26, 1995. This paper has been with the authors 19 months for 3 revisions.