SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NY 14853-3801

TECHNICAL REPORT NO. 1051

October 1993

AN EXACT ANALYSIS OF A PRODUCTION-INVENTORY STRATEGY FOR INDUSTRIAL SUPPLIERS[1]

by

Stuart A. Carr, A. Refik Güllü,
Peter L. Jackson and John A. Muckstadt

[1] Research was supported in part by the National Science Foundation grant DDM-9215328.

An Exact Analysis of a Production-Inventory Strategy for Industrial Suppliers

Stuart A. Carr
A. Refik Güllü
Peter L. Jackson
John A. Muckstadt

October, 1993

Abstract

New models of production-inventory environments are needed to reflect what many industrial suppliers face today. We propose such a model, along with a production and inventory strategy. Our strategy concentrates inventory in higher demand items, while giving lower demand items priority in production, where the partition of the items is similar to that given by Pareto analysis. We justify the strategy qualitatively and quantitatively, the latter using a simplified queuing model of the production environment. In the process, we derive recursive formulae for the probabilities of number in an $M/D/1$ system with two priority classes, preemptive-resume discipline between priority classes and FCFS discipline within a class.

Over the past 35 years, numerous models of production-inventory systems have been proposed. There are deterministic models in which all the elements, such as demand, leadtime, and rate of production, are known constants. Typically, all demand is met on time in these models. The issues addressed using these models relate to batch sizes, reorder points and where and how much inventory to hold. Solutions to these models range from the EOQ formula for the single-stage, single-item problem with constant demand, to the Wagner and Whitin [17] dynamic program for the single-stage, single-item, multi-period problem, to Muckstadt and Roundy's [9] nested power-of-two policies for the multi-stage, multi-item problem with constant demand.

Other models exist in which one or more of the elements are stochastic, and often demand is one of them. Typically, it is assumed that if demand is not met as soon as it is known (i.e. the product isn't "on the shelf"), then either that demand is lost forever or it is backordered at some additional cost. "Fill rate" often refers to the fraction of orders or units ordered that are not lost or backordered. Due to the stochastic elements, it is often impossible or economically undesirable to provide a fill rate of 100%. The randomness also makes these models much more difficult to analyze than their deterministic counterparts. Analytic solutions range from the so-called "newsboy" formula for the single-stage, single-item, single-period problem, to the (s, S) policy for the uncapacitated single-stage, single-item, infinite-horizon problem with stationary demand. Solutions to multi-stage and/or multi-item problems usually involve approximations. Both stochastic and deterministic models often fail to consider capacity limitations.

The production-inventory environments these models attempt to capture have changed. Setup times and variability in processing times have been reduced substantially in many companies. Production is more process-oriented, with "focused factories" or "factories within factories" where products flow through dedicated serial production lines. As a result, manufacturing flow times are shorter than before. Correspondingly, the applicability of most inventory theory models and their solutions has been greatly reduced, because production capacity is usually ignored in these models; that is, lead times are assumed to be independent of the quantity on order, or capacity is assumed to be infinite.

Furthermore, improved computer and communications technology has reduced the time required to receive and process orders and to deliver the finished goods to the customer. Consequently, in many instances the time between receipt of a customer order and the customer receiving the order has been reduced from months to 1–15 days. Manufacture-to-order (MTO) is now a viable option for many industrial suppliers, since their customers often order some weeks ahead of when they need their order filled.

Thus, models are needed that do not assume demand must be satisfied upon receipt of an order. And these models must be stochastic, since, under MTO, the production facility is more closely linked to the day-to-day fluctuations in customer demand.

Outline of Paper

In this paper, we propose and justify a production-inventory strategy we shall call the "No B/C Stock" strategy. The No B/C strategy concentrates inventory in the higher demand (A) items, while giving the lower demand (B/C) items priority in production, where the partition of the items resembles that of the standard A/B/C classification (an explanation of which can be found in Peterson & Silver [10]).

In Section 1, we present a model that we feel represents the environment faced by today's industrial suppliers. In Section 2, we propose a production-inventory strategy for this environment: the No B/C Stock strategy. In Section 3, we present a simpler, more tractable version of the model from Section 1. We derive exact expressions for the cost of the strategy under the model in Section 4. In Section 5, we note models from the literature that share some similarities with ours. In Section 6 we present a lower bound on the cost of the system. In Section 7, we describe our method for computing the solutions and their costs. We present our numerical results in Section 8, and summarize the paper in Section 9.

1 Our Model of the Manufacturing Environment

In this section, we describe a model of a manufacturing environment in a qualitative manner. We propose a model which we feel represents industrial suppliers in many industries. It is a composite and distillation of the various companies we have observed through years of industrial consulting.

One of the characteristics of industrial suppliers that distinguishes them from mass merchandisers is that off-the-shelf service is not always necessary. Demand arrives in the form of orders with due dates. Much of the time, customers will order a week to a month or more in advance of the time when they need the product. While the amount of advance notice depends on the industry, it tends to be greater for larger orders.

In a typical industry, a production facility consists of one or more focused factories, each of which is dedicated to a large (200+) group of products or items with similar resource and manufacturing requirements. These items fall into an A/B/C categorization according to [average] sales volume. That is to say, when the products are ranked from highest to lowest sales volume, the top 20% of the products (A items) often account for roughly 80% of total sales, the next 30% (B items) account for roughly 15% of total sales, and the remaining 50% (C items) account for some 5% of total sales. While these percentages vary by company and industry, this basic type of segmentation of products occurs in all of the many situations we have examined.

The demand patterns observed by industrial suppliers tend to be more variable than those experienced by mass merchandisers. This is due, in part, to the fact that only a few customers may account for most of the demand for an industrial item. This is particularly evident in the historical demand data of a low demand rate item where there are "spikes" corresponding to individual customer orders which are due to the customers' order/lot size rules. As a result, demand is more erratic for the lower demand rate items. In such cases, providing reasonable off-the-shelf service would require on-hand inventory to be roughly as high as the spikes. Maintaining inventory at that level — which is usually not known — would be prohibitively expensive. Not surprisingly, attempts to stock such items usually result in high holding costs and low fill rates.

The sales volume ranking (from high to low) generally matches the ranking from low to high according to relative variability (or coefficient of variation), so that the demand process for a C item is relatively more variable than that of a B item, which is relatively more variable than that of an A item. Consequently, a higher percentage of an item's average demand must be stocked for B and C items than for A items in order to achieve the same off-the-shelf fill rate. In the companies we have seen, C items, while only accounting for about 5% of demand, often account for 25% of total inventory.

Finally, as mentioned in the previous section, the length and variability of setup times and order lead times have been greatly reduced in the past few years, as well as the variability of processing times, so that MTO is becoming a more viable option.

2 The No B/C Stock Strategy

In this section we describe the No B/C Stock strategy (or "No B/C strategy", for short) by contrasting it with a more traditional production-inventory strategy. The common or traditional strategy involves holding finished goods inventories (FGI) in most, if not all, of the items and meeting demand from stock.

Under the No B/C Stock strategy, little, if any, inventory is held in some of the items, typically the B/C items (hence the name of the strategy). Instead, demand for these items is met primarily through MTO production. The strategy achieves reasonable service levels for these items by giving them priority for production. From now on, we refer to *B/C* items as the class of MTO items, which is not necessarily the same as the corresponding class resulting from the standard A/B/C analysis. Thus, inventory will be held primarily in the non-*B/C*, or *A*, items. In practice, the distinction between an *A* and a *B/C* item may not be so clear. For example, if an item experiences slow but steady demand, apart from the occasional spike, then the steady demand stream can be satisfied from stock to save on setup costs, while the spikes are met through MTO production.

The elimination of *B/C* safety stock is advantageous in many ways. Firstly, it obviates the need to calculate the appropriate safety stock levels, which is problematic for the low demand *B/C* items. Calculating safety stock levels requires the specification and parameterization of a probability distribution of demand. How is one to choose the correct distribution? There is seldom enough data to confidently reject one distribution in favor of another. So one often chooses a distribution on the basis of visually recognizing the distributional shape from a plot of demand data. This is difficult to do using the sparse and erratic historical data of a *B/C* item, as is recognizing the presence of a trend factor in the historical data that distorts the distribution.

Secondly, the No B/C strategy frees up the capital that was invested in FGI of *B/C* items and the storage space that they occupied. Thirdly, there are additional savings in holding costs beyond the cost of capital invested in inventory, since slow moving items suffer greater loss through damage and obsolescence as they spend longer on the shelf.

The main point of the strategy is to meet the challenges of the more competitive environment in which customers demand quicker response and expect lower prices. Companies can provide quick response through increased inventories and capacity, but both increase costs significantly. Instead, the No B/C strategy replaces the FGI of B/C items with an increase (hopefully smaller) in the FGI of A items. This effectively stores capacity in the form of quicker moving, and correspondingly cheaper, A item FGI. And service levels may even improve under the strategy, particularly for B/C items, which companies traditionally have had trouble stocking in the right quantities. Often, there wasn't sufficient FGI to satisfy an order, so that the customer order became a production order, and, since production was geared to a long production cycle of large batches, a small production order could have waited a long time before being filled.

We have argued qualitatively that the No B/C strategy could be better than the traditional FGI strategy, and many of the predicted benefits have been realized by the companies where this strategy has been implemented. We now seek to demonstrate the benefits of our strategy mathematically.

3 The Analytical Model

To analyze our strategy, we must specify the setting more precisely. For reasons of tractability, we will consider a less general situation than that described in Section 2. The environment we will examine will also be less favorable to our strategy in certain respects than environments we have found in practice. By performing well (see Section 8) in this more demanding environment, the strategy is shown to have a certain robustness.

We assume the production facility consists of a single constrained resource, which is realistic if a production facility's throughput is consistently limited by a single machine. The capacity limitations of the machine are captured by modeling the machine as a single-server queue, so that only one job is worked on at a time. We assume there are N items produced at this facility and that it takes a constant amount of machine time to produce one unit of any item. Furthermore, setup times are assumed to be negligible and will be ignored in the analysis.

We assume unit demands, and that an $(S_i - 1, S_i)$ control policy is used for each item i. That is, each demand is for exactly one unit of one of the items. Assuming the inventory position of item i was S_i before the demand occurred, it will be $S_i - 1$ after the demand. Thus, a production order for one unit of item i is immediately issued, bringing the inventory position back up to S_i. This makes things tractable, since the number of units of item i on the shelf is just S_i less the number of outstanding production orders for item i.

The assumptions of no setups and unit production orders are extreme examples of low setup times and small batch sizes, and they favor MTO in that there is less incentive to hold inventory, because the inventory carried does not reduce setup times or cost. Inventory is only needed to protect against the machine being busy, and the variability in the delay in obtaining the server is reduced with smaller production orders and constant production times. Other assumptions regarding the demand process could be made that do not favor the MTO strategy.

For example, we further assume orders are due upon receipt (i.e. there is no advance notice), and that they are penalized per time unit they are late. Thus, unless the item is on the shelf, some backorder cost will be incurred while the corresponding production order waits to be filled.

We assume the demands for an item arrive randomly; that is to say, the interarrival times of the orders are independent and exponentially distributed. Thus, the items' demand distributions have the desired property that the variability is inversely related to the mean (i.e. the demand for item i in one time unit is distributed as a Poisson random variable with parameter and mean λ_i, and coefficient of variation $\equiv \sigma/\mu = 1/\sqrt{\lambda_i}$, where λ_i is the average arrival rate of orders for item i). The arrival processes of orders for different items are independent. Thus, since the arrival process of production orders is the same as the arrival process of customer orders, the production facility becomes an $M/D/1$ queue.

3.1 Production Strategies

Traditionally, the processing of orders occurs on a first come, first served (FCFS) basis. The No B/C strategy gives B/C items strict production priority over the A items, so that there are two priority classes. For reasons of tractability, preemptive-resume discipline is used between classes (i.e. a B/C item will bump an A item out of service; the A item returns to service from the point it left off once all orders for B/C items have been completed) and FCFS discipline is used within a priority class. Note that FCFS within a priority class is probably inferior to a strategy that gives backorders priority over stock replenishment orders.

Furthermore, under the No B/C strategy, we require that $S_i = 0$ if i is a B/C item; that is, we do not hold *any* inventory in the B/C items. We impose this restriction to strengthen the argument for the No B/C strategy, and to avoid the counter-intuitive case of a production order for stock replenishment of a B/C item receiving service ahead of a production order to meet a backorder of an A item. Note also that the traditional strategy is just a special case of the No B/C strategy when all the items are considered as A items. Thus, the optimal No B/C strategy cannot perform worse than the traditional strategy in this model. The question is how much better can it be.

4 Cost Derivation of Analytical Model

In this section, we define the cost function for the analytical model. We derive an expression for its minimum value under $(S-1, S)$ policies. In Section 4.1, we derive the distribution of the steady-state number of outstanding production orders when there are two priority classes, which we need in order to compute the optimal value for S_i.

We will first need some notation. Let

N = total number of items.

S_i = order-up-to level of item i.

I_i = random variable denoting the steady-state on-hand inventory level of item i, where a negative number represents the number of outstanding backorders.

Q_i = random variable denoting the steady-state number of outstanding production orders for item i; that is, the number of item i's in the queue, plus one if the server is currently working on item i.

λ_i = average demand rate for item i.

h = cost of carrying one item in inventory per unit time.

p = backorder cost per item per unit time.

We use the following cost function to represent the steady-state expected cost per unit time:

$$C \equiv \sum_{i=1}^{N} \{hE[I_i^+] + pE[I_i^-]\}, \tag{1}$$

where $I_i^+ = \max\{0, I_i\}$, and $I_i^- = \max\{0, -I_i\}$. Our method will not depend on a common h and p, i.e. h_i and p_i could be used instead; however, the lower bound described in Section 6 is only applicable for a common h and p. As explained in the previous section, under an $(S_i - 1, S_i)$ policy, $I_i = S_i - Q_i$, in which case equation (1) becomes

$$\begin{aligned} C^S(S_1, \ldots, S_N) &\equiv \sum_{i=1}^{N} \{hE[(S_i - Q_i)^+] + pE[(S_i - Q_i)^-]\} \\ &= \sum_{i=1}^{N} g_i(S_i), \end{aligned} \tag{2}$$

where

$$\begin{aligned} g_i(S_i) &\equiv hE[(S_i - Q_i)^+] + pE[(S_i - Q_i)^-], \\ &= hE[(S_i - Q_i)^+] + p(EQ_i - S_i + E[(S_i - Q_i)^+]) \\ &= p(EQ_i - S_i) + (h+p)E[(S_i - Q_i)^+] \\ &= p(EQ_i - S_i) + (h+p)\sum_{k=0}^{S_i - 1}(S_i - k)P[Q_i = k], \end{aligned} \tag{3}$$

which is just the so-called newsboy problem. Since $g_i(S_i)$ is convex in S_i, a first difference argument can be used to show that S_i^* minimizes $g_i(S_i)$, where

$$S_i^* = \min\left\{ s : P[Q_i \le s] \ge \frac{p}{h+p} \right\}. \tag{4}$$

To compute S_i^*, we need to know the distribution of Q_i, which will depend on the queuing discipline used. In particular, we need to know the distribution when there are two priority classes, which is the case under the No B/C strategy. The traditional strategy involves a single priority class, which will just be a special case of our results for two priority classes.

Since $S_i = 0$ for all B/C items, the cost of the No B/C strategy is

$$\begin{aligned} C^{NoB/C}(\mathbf{S}, A) &\equiv \sum_{i \in A} g_i(S_i) + \sum_{i \in B/C} g_i(0) \\ &= \sum_{i \in A} g_i(S_i) + \sum_{i \in B/C} pEQ_i. \end{aligned} \tag{5}$$

Note that $C^{NoB/C}(\mathbf{S}, \{1, \ldots, N\}) = C^S(S_1, \ldots, S_N)$. The problem, then, is to find the partition of the items into the high and low priority classes (B/C and A, respectively), and the values $S_i^*, \forall\, i \in A$, that minimize expression (5).

4.1 Results for One and Two Priority Classes

In this section, we derive the distribution of the steady-state number in system for a multi-item FCFS $M/G/1$ system with a single priority class (which is the same as no priority classes), and a multi-item FCFS $M/D/1$ system with two priority classes with preemptive-resume discipline.

More specifically, we derive recursive expressions for $P[Q_L = k]$ and $P[Q_H = k]$, where Q_L and Q_H denote the steady-state number in system for the low and high priority classes, respectively. $P[Q_i = k]$, for an individual item i, is then expressed as a function of item i's class probabilities.

We need some additional notation. Define

$$\begin{aligned}\mathcal{L} &\equiv \text{the set of items in the low priority class.}\\ \mathcal{H} &\equiv \text{the set of items in the high priority class.}\end{aligned}$$

For now, the distinctions between the items within a priority class will be unimportant, so we define

$$\begin{aligned}\lambda_L &\equiv \textstyle\sum_{i\in\mathcal{L}}\lambda_i & Q_L &\equiv \textstyle\sum_{i\in\mathcal{L}}Q_i\\ \lambda_H &\equiv \textstyle\sum_{i\in\mathcal{H}}\lambda_i & Q_H &\equiv \textstyle\sum_{i\in\mathcal{H}}Q_i\\ \lambda &\equiv \lambda_L+\lambda_H & \rho &\equiv \rho_L+\rho_H,\end{aligned}$$

where ρ_L and ρ_H denote the traffic intensity (mean arrival rate/mean service rate) of the low and high priority classes, respectively. (Note that in the case of unit service times, $\rho_{(\cdot)} \equiv \lambda_{(\cdot)}$.)

Let us define some terminology used in the context of priority queues. The *busy period* is defined as the length of time that begins with a customer arriving into an empty system and ends the next time the system becomes empty. Hence, it is the period over which the server is continuously busy. The *completion time* of a unit is defined as the period that begins the instant service begins on a unit and ends the instant the server becomes free to take the next unit of that class. Of course, in a preemptive queuing discipline, the completion time of a lower priority item can be much higher than the unit's service time, since the lower priority item may be preempted and have to wait for several higher priority items to be serviced before completing its service.

Let $\Pi_L(\theta)$ denote the probability generating function of the stationary number of low priority items in the system. Then, by Jaiswal [8], we have, for $\rho < 1$,

$$\Pi_L(\theta) = (1-\rho)\left\{1+\frac{\lambda_H}{\lambda_L}\frac{1-b(\lambda_L(1-\theta))}{1-\theta}\right\}\left\{\frac{(1-\theta)c(\lambda_L(1-\theta))}{c(\lambda_L(1-\theta))-\theta}\right\}, \tag{6}$$

where $c(\theta)$ and $b(\theta)$ are the Laplace-Stieljes transforms (LSTs) for the <u>c</u>ompletion time of a low priority item, and the <u>b</u>usy period if the low priority items are ignored, respectively. It turns out that $c(\theta)$ satisfies (see Jaiswal [8], p.85)

$$c(\theta) = U_L(\lambda_H(1-b(\theta))+\theta),$$

and $b(\theta)$ satisfies (see Jaiswal [8], p.10)

$$b(\theta) = U_H(\lambda_H(1-b(\theta))+\theta), \tag{7}$$

where U_L and U_H are the LSTs of the service time distributions for low and high priority items, respectively. Note that when the service time distributions are identical for both priority classes, we obtain $c(\theta) = b(\theta)$. In our case of unit service times,

$$U_L(x) = U_H(x) = \int_0^\infty e^{-sx}dF(s) = e^{-x}, \tag{8}$$

where F represents the service time distribution. However, we use the general U whenever possible.

Similarly, define $\Pi_H(\theta)$ to be the probability generating function of the stationary number of high priority items in the system. It should be observed that under the preemptive-resume discipline, the low priority items do not affect the high priority items at all. Thus, $\Pi_H(\theta)$ is the probability generating function for the stationary number in system for a FCFS $M/G/1$ queue with arrival rate λ_H, traffic intensity ρ_H, and where the LST of the processing time distribution is U_H. We have (see, for example, Cohen [2]), for $\rho_H < 1$,

$$\Pi_H(\theta) = \frac{(1-\rho_H)(1-\theta)U_H(\lambda_H(1-\theta))}{U_H(\lambda_H(1-\theta))-\theta}. \tag{9}$$

We derive the probability mass functions of Q_L and Q_H from their generating functions through differentiation. That is to say, if

$$\Pi(\theta) \equiv \sum_{i=0}^{\infty} P[X=i]\,\theta^i,$$

then, for $n = 0, 1, \ldots$,

$$P[X=n] = \frac{1}{n!}\frac{d^n}{d\theta^n}\Pi(\theta)|_{\theta=0}. \tag{10}$$

(See, for example, Resnick [12]). For the sake of brevity, we shall write $f^{(k)}(s)$ for $\frac{d^k}{dx^k}f(x)|_{x=s}$. In the appendix, we derive the formulas for $\Pi_L^{(n)}(\theta)$ and $\Pi_H^{(n)}(\theta)$ that were used to prove the following propositions:

Proposition 1 *If* $\rho_H < 1$,

$$P[Q_H = 0] = 1 - \rho_H,$$

and, for $k = 1, 2, \ldots$,

$$\begin{aligned} P[Q_H = k] = {} & U_H(\lambda_H)^{-1} \\ & \cdot \Big(P[Q_H = k-1] \\ & -\sum_{j=1}^{k}\frac{1}{j!}(-\lambda_H)^j U_H^{(j)}(\lambda_H)\,P[Q_H = k-j] \\ & +\frac{(1-\rho_H)}{k!}\left[(-\lambda_H)^k U_H^{(k)}(\lambda_H) - k(-\lambda_H)^{k-1}U_H^{(k-1)}(\lambda_H)\right]\Big). \end{aligned} \tag{11}$$

The above proposition holds for any $M/G/1$ queue with FCFS service, arrival rate λ_H, traffic intensity ρ_H, and LST of service time distribution U_H. However, the next proposition only holds for an $M/D/1$ queue with unit service times, i.e. $U_L(x) = U_H(x) = e^{-x}$.

Proposition 2 *If $\rho < 1$,*

$$P[Q_L = 0] = (1-\rho)\left(1 + \frac{\lambda_H}{\lambda_L}\left[1 - b(\lambda_L)\right]\right), \tag{12}$$

and, for $k = 1, 2, \ldots$,

$$\begin{aligned} P[Q_L = k] &= b(\lambda_L)^{-1} \\ &\cdot \Big(P[Q_L = k-1] \\ &- \sum_{j=0}^{k-1} \frac{1}{(k-j)!} (-\lambda_L)^{k-j}\, b^{(k-j)}(\lambda_L)\, P[Q_L = j] \\ &+ \frac{1}{k!}(1-\rho)(-\lambda_L)^{k-1}\left[(2-k)b^{(k-1)}(\lambda_L) + (2-\lambda_L-\lambda_H)b^{(k)}(\lambda_L)\right]\Big), \end{aligned} \tag{13}$$

where, for $k \geq 1$,

$$b^{(k)}(\theta) = \frac{-b^{(k-1)}(\theta) + \lambda_H \sum_{j=1}^{k-1}\binom{k-1}{j} b^{(j)}(\theta) b^{(k-j)}(\theta)}{1 - \lambda_H b(\theta)}. \tag{14}$$

Proof: The proofs for the propositions can be found in the Appendix. They involve straightforward but messy algebra to establish, by induction, recursive formulas for the derivatives of the generating functions Π_L and Π_H. □

Proposition 3

$$P[Q_i = k] = \sum_{m=k}^{\infty} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_L}\right)^k \left(1 - \frac{\lambda_i}{\lambda_L}\right)^{m-k} P[Q_L = m], \tag{15}$$

for $i \in \mathcal{L}$. The same result is valid for any $i \in \mathcal{H}$ with subscripts L replaced by H.

Proof: Fix $i \in \mathcal{L}$.

$$\begin{aligned} P[Q_i = k] &= \sum_{m=k}^{\infty} P[Q_i = k, Q_L = m] \\ &= \sum_{m=k}^{\infty} P[Q_i = k | Q_L = m]\, P[Q_L = m]. \end{aligned}$$

Since the arrivals to the queuing system follow a Poisson process, and the customers within the same priority class are treated on a FCFS basis, the [steady-state] probability that any low priority item in the system is of type i is λ_i/λ_L. Let $X_j = 1$ if the j^{th} low priority item in the system is of type i, where the items are ordered by their position in the low priority queue, the head of which is the item in service if that item belongs to $\mathcal{L}$. Otherwise, let $X_j = 0$. Then $Q_i = X_1 + X_2 + \cdots + X_{Q_L}$. Conditioned on Q_L, the X_j's are i.i.d. Bernoulli random variables with mean λ_i/λ_L, and so $Q_i|Q_L \sim$ Binomial$(Q_L, \lambda_i/\lambda_L)$. Thus,

$$P[Q_i = k|Q_L = m] = \binom{m}{k}\left(\frac{\lambda_i}{\lambda_L}\right)^k\left(1-\frac{\lambda_i}{\lambda_L}\right)^{m-k}. \ \square$$

5 Literature Review

The integrated analysis of inventory and queuing models is not new (the reader is referred to Prabhu [11] for a discussion of earlier models). We will just mention here a few recent works that involve models similar to ours.

Zipkin [19] formulates a multi-item production problem in which the production facility is represented by a network of queues. Through various approximations, he formulates a convex program to determine optimal batch sizes. Zheng and Zipkin [20] consider a production-inventory problem involving two items in which the production facility is represented by an $M/M/1$ queue. They compare the performances of order-up-to S policies under different priority rules: FCFS and Longest Queue First Served. They show that the expected costs of holding and backorders are lower for the LQFS discipline.

Perhaps closest to the spirit of our model is Williams [18], in which he considers a multi-item production-inventory problem which he models as a multi-server, non-preemptive, delay- and class-dependent priority queuing problem. He partitions the items into two classes: Make-to-Stock (MTS) and MTO. The MTS items follow (Q,r) policies. The MTO items have a different cost function which penalizes backorders only when they have been backordered beyond a certain amount of time. There are several differences in our approach. By penalizing backorders immediately, we can use the same cost function for an item independent of whether it is MTS or MTO. This allows meaningful cost comparisons across different MTS/MTO partitions. By considering a simpler system, we are able to obtain exact expressions for the cost of the system. Williams' model contains several desirable features that our analytical model does not; however, its complexity results in the necessity for making approximations.

The lower bound we describe in Section 6 involves a single-item $M/D/1$ queue. For a treatment of the finite capacity, single-item production/inventory problem in discrete time, and for the other references, we refer the reader to Federgruen and Zipkin [3] & [4], and Tayur [16]. For finite production rate problems in continuous time, Güllü and Jackson [7] provide the related literature and some new results. Heyman [5] investigates optimal operating policies for $M/G/1$ queuing systems with server start-up and shut-down costs, and a cost per unit time spent in the system for each customer. Sobel [15] considers a $GI/G/1$ queuing system operating under a very general cost structure. He shows that almost any pure stationary policy is equal to that of an (M, m) policy: if the queue length is less than or equal to m, then do not provide service until it increases to M (where $M > m$), at which point service begins and continues until the queue length drops to m again. An application of the (M, m) policy on an $M/D/1$ queuing system integrated with an inventory model is performed by Gavish and Graves [6]. When there are no start-up and shut-down costs for the server, this policy corresponds to an order-up-to S policy for the production-inventory problem.

6 The Lower Bound

The purpose of this section is to present a lower bound on the expected cost function, (1), of the multi-item production-inventory system described in Section 3. In particular, this lower bound applies to the cost of an optimal policy for the multi-item system, which we denote by C^O. Note that the optimal policy may not necessarily involve priority classes or $(S-1, S)$ policies.

Given an instance of the problem described in Section 3, a lower bound can be derived from the following single item problem. The production facility is the same (with unit processing times), and the demand process for the single item is given by the superposition of the demand processes of all the items present in the original system. The optimal policy of the single item system is a lower bound for any policy in the multi-item system, since any policy from the latter can be implemented in the former by randomly (but with the appropriate distribution) labeling demands $1, \ldots, N$, and labeling inventory and production orders according to the multi-item policy.

The optimal policy for the single-item system is just an $(S-1, S)$ policy (see Gavish and Graves [6]), in which case the production facility becomes an $M/D/1$ queue with an arrival rate and traffic intensity of λ. The cost of this system is:

$$C^{single\ item}(S) \equiv hE[I^+]+pE[I^-] = hE[(S-Q^{single\ item})^+]+pE[(S-Q^{single\ item})^-], \quad (16)$$

where $Q^{single\ item}$ denotes the steady-state number of outstanding production orders. It is minimized as expression (3) was, namely by

$$S^* = \min\left\{s : P[Q^{single\ item} \leq s] \geq \frac{p}{h+p}\right\}.$$

The distribution of $Q^{single\ item}$ can be computed using Proposition 1, where $\lambda_H = \lambda$, $\rho_H = \rho$, and $U_H(x) = e^{-x}$.

Thus, we have

$$C \geq C^O \geq C^{single\ item}(S^*),$$

which means that if C under the No B/C strategy is close to $C^{single\ item}(S^*)$, then the No B/C Strategy is close to optimal.

7 Computational Method

In this section, we describe how we computed the cost of the No B/C strategy given by expression (5), and how we tried to minimize it. There are two parameters over which to minimize: the partition of the items into the priority classes, A and B/C, and the vector $\mathbf{S}$ of order-up-to levels for the A items.

7.1 Partitioning the Items

Minimization with respect to the partition was done by brute force enumeration. We looked at various (A, B/C) partitions of the items and calculated the optimal cost under each partition. We then took the lowest cost partition as our answer.

To ensure our answer minimized expression (5), we would have to consider all 2^N possible partitions. This number is prohibitively large for values of N in our region of interest, i.e. $N \geq 200$. So we only considered partitions that were simple splits. That is to say, we ranked the items by demand rate and only considered partitions of the form where the A class consisted of the n highest demanded items, and the B/C class of the remaining $N - n$ lower demand items.

We chose to consider simple splits because we expected them to contain an optimal or near-optimal solution to expression (5). We found that to be the case for small values of N for which we could examine all possible subsets. When considering 15 values of ρ between 0.5 and 1, with various demand profiles for 2, 3, 4, and 5 items, and with $h = 1$ and $p = 2, 10, 20$, we found that a simple split was optimal in all cases.

Unfortunately, we have not been able to prove the optimality of simple splits, nor an error bound resulting from the restriction to simple splits.

7.2 Computing the Cost

To find the value of expression (5), we need to compute $g_i(S_i)$ for all the A items, and EQ_i for all the items. The latter is computed using the following lemma:

Lemma 4

$$EQ_i = \frac{\lambda_i}{\lambda_A} EQ_A, \forall\, i \in A. \tag{17}$$

The same result holds with A replaced by B/C.

Proof: Without loss of generality, fix $i \in A$.

$$EQ_i \;=\; \sum_{k=1}^{\infty} kP[Q_i = k] \tag{18}$$

$$
\begin{aligned}
&= \sum_{k=1}^{\infty} k \sum_{m=k}^{\infty} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1-\frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A=m] && (19)\\
&= \sum_{m=1}^{\infty} \sum_{k=1}^{m} k \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1-\frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A=m] && (20)\\
&= \sum_{m=1}^{\infty} P[Q_A=m] \sum_{k=1}^{m} k \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1-\frac{\lambda_i}{\lambda_A}\right)^{m-k} && (21)\\
&= \sum_{m=1}^{\infty} P[Q_A=m] m \frac{\lambda_i}{\lambda_A} && (22)\\
&= \frac{\lambda_i}{\lambda_A} \sum_{m=1}^{\infty} m P[Q_A=m] && (23)\\
&= \frac{\lambda_i}{\lambda_A} EQ_A, && (24)
\end{aligned}
$$

where equation (19) follows from equation (15), and equation (20) from Tonelli's Theorem. The second sum in equation (21) is just the expected value of a binomial random variable with mean $m\lambda_i/\lambda_A$. □

Jaiswal [8], p. 96, gives formulas for EQ_A and $EQ_{B/C}$. In our case of unit service times, they reduce to:

$$
\begin{aligned}
EQ_A &= \frac{1}{1-\rho_{B/C}} \left[\lambda_A + \frac{1}{2} \cdot \frac{\lambda_A \lambda_{B/C} + \lambda_A^2}{1-\rho}\right],\\
EQ_{B/C} &= \rho_{B/C} + \frac{1}{2} \cdot \frac{\lambda_{B/C}^2}{1-\rho_{B/C}}.
\end{aligned}
$$

Given a partition of the items into the two priority classes, expression (5) is minimized by setting $S_i = S_i^*$ for all the A items, where S_i^* is given by equation (4). We need to know only the first few values of the distribution of Q_i in order to compute S_i^* and then to calculate $g_i(S_i^*)$ as given by equation (3). Equation (15) gives a formula for $P[Q_i = k]$, but it involves an *infinite* sum of $P[Q_A = m]$ terms. We use Proposition 2 to calculate $P[Q_A = m]$ in terms of $P[Q_A = 0], P[Q_A = 1], \ldots, P[Q_A = m-1]$. Since we can only calculate a finite number of these terms, we approximate $P[Q_i = k]$ by fitting a tail with geometric decay to the distribution of Q_A:

$$
\begin{aligned}
P^{approx}[Q_i=k] \equiv\ & \sum_{m=k}^{T} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1-\frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A=m] +\\
& \sum_{m=T+1}^{\infty} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1-\frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A=T]\, r^{m-T} && (25)\\
=\ & \sum_{m=k}^{T-1} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1-\frac{\lambda_i}{\lambda_A}\right)^{m-k}
\end{aligned}
$$

$$\cdot\left\{P[Q_A = m] - P[Q_A = T]\, r^{m-T}\right\}$$
$$+ \frac{P[Q_A = T]}{r^T} \frac{1}{1 - r(1 - \frac{\lambda_i}{\lambda_A})} \left(\frac{\frac{\lambda_i}{\lambda_A}}{r^{-1} - 1 + \frac{\lambda_i}{\lambda_A}} \right)^k, \qquad (26)$$

where $T \geq k$, and $P[Q_A = 0], \ldots, P[Q_A = T]$ are known, and $r \equiv \left(1 + \frac{P[Q_A = T]}{1 - P[Q_A \leq T]}\right)^{-1}$ so that these tail probabilities complete the distribution: i.e. $\sum_{m=0}^{T} P[Q_A = m] + \sum_{m=T+1}^{\infty} P[Q_A = T]\, r^{m-T} = 1$. (See the appendix for the derivation of the last equality.) When evaluating the quality of this approximation, we observed that the ratio of successive probabilities of number in system for both the A and B/C classes quickly converged to the value used as the ratio in the fitted tail, although the speed of convergence appeared to be inversely related to ρ. Further empirical evidence of the accuracy of this approximation was that the computed cost was insensitive to changes in T for $T \geq 5$.

In order to get a sense of how accurate our approximation was to $P[Q_i = k]$, we also computed

$$P^{LB}[Q_i = k] \equiv \sum_{m=k}^{T} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1 - \frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A = m] \qquad (27)$$

and

$$\begin{aligned} P^{UB}[Q_i = k] &\equiv \sum_{m=k}^{T} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1 - \frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A = m] + \\ &\quad \sum_{m=T+1}^{\infty} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1 - \frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A = T] \\ &= \sum_{m=k}^{T-1} \binom{m}{k} \left(\frac{\lambda_i}{\lambda_A}\right)^k \left(1 - \frac{\lambda_i}{\lambda_A}\right)^{m-k} \{P[Q_A = m] - P[Q_A = T]\} \\ &\quad + \frac{P[Q_A = T]}{\frac{\lambda_i}{\lambda_A}}. \end{aligned} \qquad (28)$$

(See the appendix for the derivation of the last equation). P^{LB} is simply equation (15) with the sum truncated at T, while P^{UB} involves replacing the tail probabilities (i.e. $P[Q_A = m]$, $m > T$) with $P[Q_A = T]$. Thus, provided T is sufficiently large so that $P[Q_A = m] \leq P[Q_A = T]$ for all $m \geq T$,

$$P^{LB}[Q_A = k] \leq P[Q_A = k] \leq P^{UB}[Q_A = k]. \qquad (29)$$

Note also that

$$P^{LB}[Q_A = k] \leq P^{approx}[Q_A = k] \leq P^{UB}[Q_A = k]. \qquad (30)$$

Equation (29) implies that

$$\underline{S}_i \leq S_i^* \leq \overline{S}_i, \tag{31}$$

where S_i^* is given by equation (4), and $\underline{S}_i$ and $\overline{S}_i$ are given by equation (4) with P replaced by P^{UB} and P^{LB}, respectively. We would also like to replace P in equation (3); the following definition replaces P with P^t:

$$g_i^t(S_i) \equiv p(EQ_i - S_i) + (h+p)\sum_{k=0}^{S_i-1}(S_i - k)P^t[Q_i = k].$$

Then,

$$\min_{\underline{S}_i \leq S_i \leq \overline{S}_i} g_i^{LB}(S_i) \leq g_i^{LB}(S_i^*) \leq g_i(S_i^*) \leq g_i^{UB}(S_i^*) \leq \max_{\underline{S}_i \leq S_i \leq \overline{S}_i} g_i^{UB}(S_i). \tag{32}$$

The outer inequalities follow from inequalities (31), while the inner inequalities follow from inequalities (29).

Let us define several cost functions:

$$\begin{aligned}
C^{LB}(n) &\equiv \sum_{i=1}^{n} \min_{\underline{S}_i \leq S_i \leq \overline{S}_i} g_i^{LB}(S_i) + \sum_{i=n+1}^{N} pEQ_i, \\
C^{UB}(n) &\equiv \sum_{i=1}^{n} \max_{\underline{S}_i \leq S_i \leq \overline{S}_i} g_i^{UB}(S_i) + \sum_{i=n+1}^{N} pEQ_i, \\
C^{NoB/C}(n) &\equiv \sum_{i=1}^{n} g_i(S_i^*) + \sum_{i=n+1}^{N} pEQ_i, \qquad (33) \\
\tilde{C}^{NoB/C}(n) &\equiv \sum_{i=1}^{n} g_i^{approx}(\tilde{S}_i^*) + \sum_{i=n+1}^{N} pEQ_i, \qquad (34)
\end{aligned}$$

where $\tilde{S}_i^*$ is given by equation (4) with P^{approx} in place of P. Note that $C^{NoB/C}(n) = \min_{\mathbf{S}} C^{NoB/C}(\mathbf{S}, \{1, \ldots, n\})$, which is to say that $C^{NoB/C}(n)$ gives the minimum cost of expression (5) for fixed A.

It then follows from inequalities (32) that

$$C^{LB}(n) \leq C^{NoB/C}(n) \leq C^{UB}(n). \tag{35}$$

Inequalities (31) also hold with $\tilde{S}_i^*$ in place of S_i^* (from inequalities (30)), so then the relation analogous to inequalities (32) also holds:

$$\min_{\underline{S}_i \leq S_i \leq \overline{S}_i} g_i^{LB}(S_i) \leq g_i^{LB}(\tilde{S}_i^*) \leq g_i^{approx}(\tilde{S}_i^*) \leq g_i^{UB}(\tilde{S}_i^*) \leq \max_{\underline{S}_i \leq S_i \leq \overline{S}_i} g_i^{UB}(S_i).$$

Thus,

$$C^{LB}(n) \leq \tilde{C}^{NoB/C}(n) \leq C^{UB}(n). \tag{36}$$

We found we could make C^{LB} and C^{UB} arbitrarily close by increasing the cutoff point, T. Hence, by inequalities (35) and (36), $\tilde{C}^{NoB/C}$ was arbitrarily close to $C^{NoB/C}$, so our approximation was essentially exact to within the precision of the machine.

Thus, when the cost of the No B/C strategy is quoted, we are using $\tilde{C}^{NoB/C}$. Figure 1 gives some examples of $\tilde{C}^{NoB/C}$.

7.3 Summary of Computational Method

We compute $C^{single\ item}(S^*)$, the lower bound, and $\tilde{C}^{NoB/C}(n)$, for $n = 1, \ldots, N$, the minimum of which we take as the cost of the No B/C strategy. (Note that $\tilde{C}^{NoB/C}(0) \geq \tilde{C}^{NoB/C}(N)$, since $\mathbf{S} = 0$ is a possible solution to the latter. Thus, the $n = 0$ case need not be considered.)

We calculate these costs by first computing the first few class probabilities. When there is only a single priority class (which occurs in the single item situation and when $n = N$), we use the formulas from Proposition 1. When there are two priority classes (i.e. $n = 1, \ldots, N-1$), the formulas from Proposition 2 are used to compute the class A probabilities. With both formulas, logarithms are used to control the size of the factorial terms. Logarithms are also used to evaluate $b^{(k)}(\lambda_L)$, since it would increase roughly one thousand fold in absolute value with each higher derivative. $b(\lambda_L)$ is determined from equation (7) numerically, using the bisection method.

For each A item i, P^{approx} is computed using equation (26), where the cutoff point, T, is determined experimentally so that the cost is unchanged by further increases in T. Then $\tilde{S}_i^*$, $g_i^{approx}(\tilde{S}_i^*)$, and finally $\tilde{C}^{NoB/C}(n)$ are computed.

The same formulae were used to calculate each item's cost in the case of a single priority class; only the computation of the class probabilities changed.

8 Results

In this section, we present numerical results for several cases and make some inferences based on those results. The analytical model is parameterized by the number of items (N), their arrival rates (λ_i), and the holding and backorder costs (h and p).

Once again, we index the items from highest to lowest demand rate. We require the capacity utilization or traffic intensity, ρ, to be less than 1 so that the steady-state is meaningful and our formulas apply. Since we have unit service times,

$$1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > 0$$

and

$$\lambda_1 + \lambda_2 + \cdots + \lambda_N = \rho.$$

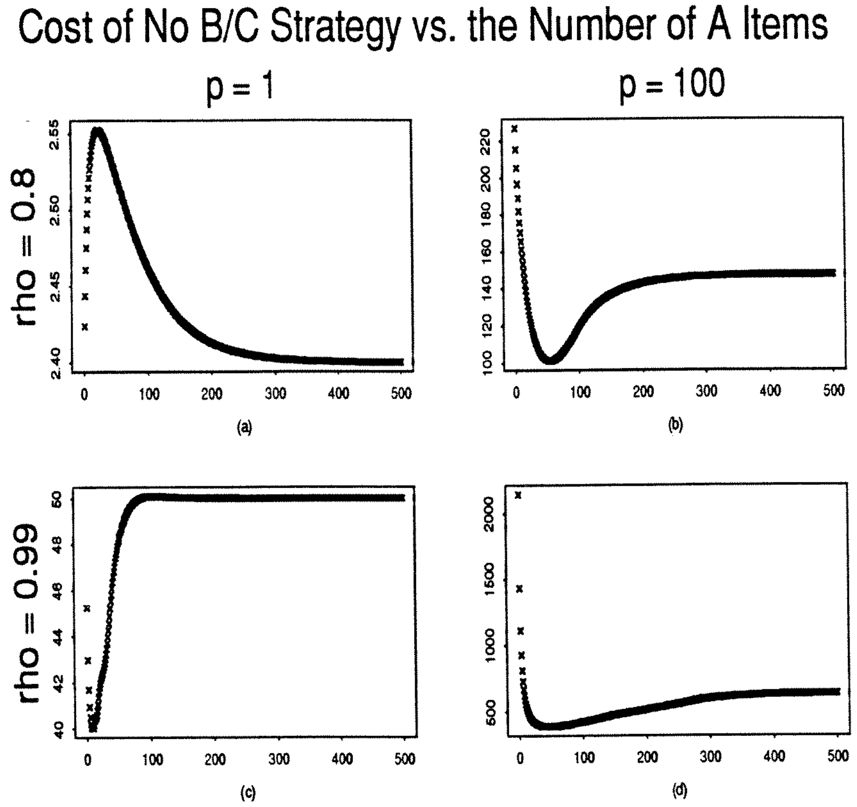

Figure 1: The above graphs give the cost of the No B/C strategy as a function of the number of A items, n, where A always consists of the n highest demanded items (i.e. $A = \{1, \ldots, n\}$). In all the cases, there were 500 items ($N = 500$) and unit holding costs ($h = 1$). This figure illustrates the three basic shapes we observed for the cost function $\widetilde{C}^{NoB/C}$. The graph in (a) is monotone increasing, then monotone decreasing. In this case, the minimum split was to have all the items in the A class, which is just the FCFS solution. In some cases when we have this shape of graph, the first point has been the minimum: i.e. $A = \{1\}$. The graphs in (b) and (d) are monotone decreasing, then monotone increasing so that there is a local minimum. This is the more "typical" case we observed. The graph in (c) is monotone decreasing, then monotone increasing, then monotone decreasing again, so that it has both a local maximum and a local minimum.

		No B/C Strategy			FCFS Strategy		
	Lower Bound	Cost		Total	Total	Cost	
ρ	on Cost	(1)	$\lvert A\rvert$	Inventory	Inventory	(2)	(1)/(2)
0.6	$2.72	$10.50	100%	0	0	$10.50	100%
0.65	$3.09	$12.54	100%	0	0	$12.54	100%
0.7	$3.68	$14.97	6%	2	0	$15.17	99%
0.75	$4.51	$17.48	12%	4	0	$18.75	93%
0.8	$5.65	$20.22	20%	7	0	$24.00	84%
0.85	$7.64	$23.81	28%	10	2	$32.52	73%
0.9	$11.62	$30.21	38%	15	15	$45.77	66%
0.95	$23.59	$45.83	36%	28	36	$72.73	63%
0.99	$119.50	$145.49	32%	126	156	$215.75	67%
0.995	$239.39	$265.96	32%	245	290	$358.99	74%
0.999	$1198.55	$1225.60	32%	1204	1281	$1374.96	89%

Table 1: N=250, h=1, p=10

So that the demand pattern follows the 80/20 rule of the standard A/B/C classification described in Section 1, we require that

$$\lambda_1 + \lambda_2 + \cdots + \lambda_{\lfloor 0.2N \rfloor} = 0.8\rho.$$

We wanted to automatically generate the items' individual arrival rates given N and ρ. So we added the following additional constraint to determine uniquely the arrival rates:

$$\frac{\lambda_{i+1}}{\lambda_i} = \frac{\lambda_2}{\lambda_1} \quad \forall\, i = 1, 2, \ldots, N-1.$$

Tables 1–4 present the results for 250 and 500 items at various traffic intensities with two different backorder costs. Observe that the absolute difference between the cost of the No B/C strategy and the lower bound seems relatively insensitive to changes in the traffic intensity, which means that the *relative* difference gets smaller as the traffic intensity increases (along with the cost). Thus, for high traffic intensities, the No B/C strategy appears to be a "near-optimal" strategy.

The cost advantage of the No B/C strategy over the FCFS strategy improves with higher traffic intensities, although the improvement begins to diminish at the highest traffic intensities. The maximum advantage seems to occur around $\rho = 0.95$ where the No B/C strategy cost is some 60% of the cost of the FCFS strategy.

As the traffic intensity increases, the maximum inventory level (i.e. the sum of the order-up-to levels) tends to increase more slowly under No B/C than FCFS, although the No B/C strategy sometimes carries more inventory than FCFS at low traffic intensities. The percent of total demand accounted for by the A class under the No B/C strategy (indicated by $|A|$) seems to increase somewhat with increasing traffic intensity, but then begins to trail off for the highest traffic intensities.

		No B/C Strategy			FCFS Strategy		
	Lower Bound	Cost		Total	Total	Cost	
ρ	on Cost	(1)	$\|A\|$	Inventory	Inventory	(2)	(1)/(2)
0.6	$4.33	$39.42	42%	17	17	$47.32	83%
0.65	$4.99	$42.30	46%	19	22	$52.90	80%
0.7	$5.92	$45.29	51%	22	28	$59.00	77%
0.75	$7.26	$48.68	55%	25	34	$66.00	74%
0.8	$9.21	$53.09	59%	28	42	$74.61	71%
0.85	$12.50	$60.23	64%	32	52	$86.51	70%
0.9	$19.02	$74.02	67%	43	65	$107.16	69%
0.95	$38.67	$101.27	59%	66	106	$155.19	65%
0.99	$195.94	$269.63	54%	226	322	$392.11	69%
0.995	$392.53	$468.19	52%	423	547	$626.97	75%
0.999	$1965.26	$2042.56	51%	1994	2193	$2289.08	89%

Table 2: N=250, h=1, p=50

		No B/C Strategy			FCFS Strategy		
	Lower Bound	Cost		Total	Total	Cost	
ρ	on Cost	(1)	$\|A\|$	Inventory	Inventory	(2)	(1)/(2)
0.6	$2.72	$10.50	100%	0	0	$10.50	100%
0.65	$3.09	$12.54	100%	0	0	$12.54	100%
0.7	$3.68	$15.17	100%	0	0	$15.17	100%
0.75	$4.51	$18.75	100%	0	0	$18.75	100%
0.8	$5.65	$23.21	6%	4	0	$24.00	97%
0.85	$7.64	$28.20	13%	9	0	$32.58	87%
0.9	$11.62	$35.33	21%	15	0	$49.50	71%
0.95	$23.59	$53.30	21%	30	30	$92.27	58%
0.99	$119.50	$156.02	24%	128	168	$269.11	58%
0.995	$239.39	$277.04	24%	248	314	$432.74	64%
0.999	$1198.55	$1237.15	23%	1207	1337	$1503.18	82%

Table 3: N=500, h=1, p=10

		No B/C Strategy			FCFS Strategy		
	Lower Bound	Cost		Total	Total	Cost	
ρ	on Cost	(1)	$\|A\|$	Inventory	Inventory	(2)	(1)/(2)
0.6	\$4.33	\$49.01	18%	12	0	\$52.50	93%
0.65	\$4.99	\$54.21	24%	17	1	\$62.67	86%
0.7	\$5.92	\$59.23	30%	22	13	\$74.46	80%
0.75	\$7.26	\$64.33	35%	27	26	\$87.65	73%
0.8	\$9.21	\$69.99	40%	32	41	\$103.20	68%
0.85	\$12.50	\$77.50	46%	38	60	\$123.00	63%
0.9	\$19.02	\$91.59	53%	47	86	\$152.19	60%
0.95	\$38.67	\$125.62	44%	72	130	\$216.96	58%
0.99	\$195.94	\$301.11	41%	236	395	\$517.17	58%
0.995	\$392.53	\$501.19	40%	435	645	\$786.61	64%
0.999	\$1965.26	\$2076.65	39%	2005	2356	\$2539.06	82%

Table 4: N=500, h=1, p=50

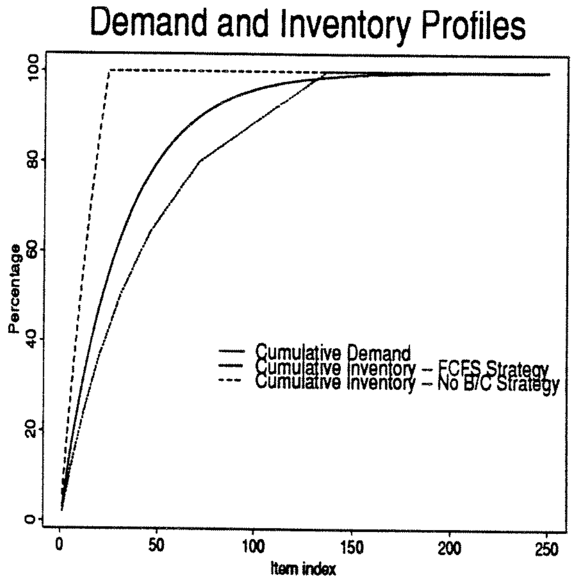

Figure 2: These graphs should be read in the following way. The cumulative demand graph is 80% at 50, which means that the 50 highest demanded items account for 80% of total demand. The cumulative inventory graphs have a similar interpretation, where inventory is measured by the order-up-to levels. For example, the cumulative inventory graph for the No B/C Strategy hits 100% at 24, which means that inventory is only carried in the first 24 items (i.e. $S_i > 0$ for $i = 1, \ldots, 24$, and $S_i = 0$ for the remaining items, $i = 25, \ldots, N$). The parameters in this case were $N = 250$, $\rho = 0.99$, $h = 1$ and $p = 50$.

Figure 2 highlights how the inventory profile differs between the No B/C strategy and the FCFS strategy. Under FCFS, the inventory profile mirrors the demand profile, since each item is stocked independently. Under the No B/C strategy, inventory is concentrated in the high demand items.

9 Summary

In this paper, we have described a production and inventory strategy that we feel is a good and quite general approach in the context that many industrial suppliers face. While we are just beginning our evaluation of this strategy, the basic idea of giving production priority to the make-to-order, low demand items while concentrating inventory in the high demand items has worked well in certain applications.

To examine the strategy more rigorously, we created a model of a manufacturing environment that was tractable and also somewhat disadvantageous to our strategy. We were able to compute the exact cost under this model, as well as a lower bound on the cost of the (unknown) optimal strategy. For higher traffic intensities, the No B/C strategy was significantly cheaper than the more traditional FCFS strategy, and its cost was relatively close to the lower bound. We feel this demonstrates the utility of the No B/C strategy.

10 Future Research

The analytical model could be extended in several ways. Other service time distributions should be considered, as well as different arrival processes, such as batch arrivals. We ultimately want to study our more realistic model, which includes setups and due dates. As such models are likely to be intractable analytically, we intend to perform simulation experiments.

11 Acknowledgments

We would like to thank Professors N. Prabhu and S. Resnick for their help on some queuing-related questions. Also, we would like to thank Dr. Frank Chance for providing us with and helping us use his Delphi [1] simulation program to check our computations. We also used Sigma [14] to validate our program.

This research was supported in part by NSF grant DDM-9215328 and by the IBM Manufacturing Research Graduate Fellowship.

A Proofs of the Propositions

A.1 Proofs of Equations (26) and (28)

The second sum in equation (25) can be written in terms of sums beginning at k: i.e. $\sum_{m=T+1}^{\infty} = \sum_{m=k}^{\infty} - \sum_{m=k}^{T}$. The $\sum_{m=k}^{T}$ term, when combined with the first sum in equation (25), gives the first sum of equation (26). Note that the upper limit of the sum is now $T-1$, because the term for $m=T$ is zero. It remains to show that the $\sum_{m=k}^{\infty}$ term equals the last term of equation (26):

$$\begin{aligned}
&\sum_{m=k}^{\infty}\binom{m}{k}\left(\frac{\lambda_i}{\lambda_A}\right)^k\left(1-\frac{\lambda_i}{\lambda_A}\right)^{m-k} P[Q_A=T]r^{m-T} \\
&= \frac{\left(\frac{\lambda_i}{\lambda_A}\right)^k P[Q_A=T]}{r^{T-k}k!}\sum_{m=k}^{\infty} m(m-1)\cdots(m-k+1)\left[r\left(1-\frac{\lambda_i}{\lambda_A}\right)\right]^{m-k} \\
&= \frac{\left(\frac{\lambda_i}{\lambda_A}\right)^k P[Q_A=T]}{r^{T-k}k!}\frac{k!}{\left[1-r\left(1-\frac{\lambda_i}{\lambda_A}\right)\right]^{k+1}} \\
&= \frac{P[Q_A=T]}{r^T}\frac{1}{1-r(1-\frac{\lambda_i}{\lambda_A})}\left(\frac{\frac{\lambda_i}{\lambda_A}}{r^{-1}-1+\frac{\lambda_i}{\lambda_A}}\right)^k .
\end{aligned}$$

The middle equality is an application of the identity

$$\sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)a^{n-k} = \frac{k!}{(1-a)^{k+1}},\ k=0,1,2,\ldots \qquad (37)$$

which follows from the Corollary on p. 173 of Rudin [13] for the power series $\sum_{n=0}^{\infty} a^n = \frac{1}{1-a}$, for $|a|<1$. Note that r and $\frac{\lambda_i}{\lambda_A}$ are both less than 1.

Equation (28) is derived in exactly the same way (i.e. by applying equation (37) again).

A.2 Proof of Proposition 2

A.2.1 Proof of (14)

Note that $b(\theta)$ satisfies

$$\begin{aligned}
b(\theta) &= U_H(\lambda_H(1-b(\theta))+\theta) \\
&= e^{-(\lambda_H(1-b(\theta))+\theta)}.
\end{aligned}$$

The first equality is just equation (7), and the second is due to the unit processing times which result in $U_H(x) = e^{-x}$. Clearly, $b(\theta) > 0$ for all θ since $e^x > 0$ for all x. Taking logarithms of both sides yields

$$ln(b(\theta)) = -(\lambda_H(1 - b(\theta)) + \theta).$$

Taking derivatives of both sides with respect to θ yields

$$b^{(1)}(\theta) \quad = \quad \frac{-b(\theta)}{1 - \lambda_H b(\theta)}, \tag{38}$$

which proves equation (14) for $k = 1$. Assume equation (14) is true for some $k \geq 1$, and thus

$$(1 - \lambda_H b(\theta))b^{(k)}(\theta) = -b^{(k-1)}(\theta) + \lambda_H \sum_{j=1}^{k-1} \binom{k-1}{j} b^{(j)}(\theta) b^{(k-j)}(\theta).$$

Differentiating both sides with respect to θ yields

$$\begin{aligned}
&(1 - \lambda_H b(\theta))b^{(k+1)}(\theta) - \lambda_H b^{(1)}(\theta) b^{(k)}(\theta) \\
&= \quad -b^{(k)}(\theta) + \lambda_H \sum_{j=1}^{k-1} \binom{k-1}{j} \{b^{(j+1)}(\theta) b^{(k-j)}(\theta) + b^{(j)}(\theta) b^{(k-j+1)}(\theta)\} \\
&= \quad -b^{(k)}(\theta) + \lambda_H \sum_{j=2}^{k-1} \left\{ \binom{k-1}{j-1} + \binom{k-1}{j} \right\} b^{(j)}(\theta) b^{(k-j+1)}(\theta) + \\
&\qquad \lambda_H \left\{ \binom{k-1}{k-1} + \binom{k-1}{1} \right\} b^{(1)}(\theta) b^{(k)}(\theta).
\end{aligned}$$

The last equality follows by writing the sum as two sums and replacing the index j of the first sum with $j - 1$, then recombining the sums. Using the identity

$$\binom{k-1}{j-1} + \binom{k-1}{j} = \binom{k}{j}, \tag{39}$$

the previous equation can be rewritten as

$$\begin{aligned}
&(1 - \lambda_H b(\theta))b^{(k+1)}(\theta) = -b^{(k)}(\theta) + \\
&\qquad \lambda_H \left\{ k b^{(1)}(\theta) b^{(k)}(\theta) + \sum_{j=2}^{k-1} \binom{k}{j} b^{(j)}(\theta) b^{(k+1-j)}(\theta) + b^{(1)}(\theta) b^{(k)}(\theta) \right\},
\end{aligned}$$

which is just equation (14) with $k+1$ in place of k and some terms slightly rearranged. Thus, by induction, equation (14) holds for all $k \geq 1$. □

Equation (14) can be made computationally more efficient by grouping like terms: i.e. grouping $b^{(j)}(\theta) b^{(k-j)}(\theta)$ and $b^{(k-j)}(\theta) b^{(j)}(\theta)$ terms. That, along with the identity

$\binom{n}{k} = \binom{n}{n-k}$ and equation (39), yields

$$
\begin{aligned}
&b^{(k)}(\theta) = (1 - \lambda_H b(\theta))^{-1} \cdot \\
&\quad \Big(- b^{(k-1)}(\theta) + \textstyle\sum_{j=1}^{\lfloor \frac{k-1}{2} \rfloor} \binom{k}{j} b^{(j)}(\theta) b^{(k-j)}(\theta) + 1_{\{\text{k is even}\}} \binom{k-1}{\frac{k}{2}} \left[b^{(\frac{k}{2})}(\theta) \right]^2 \Big),
\end{aligned}
$$

where $1_{\{\text{k is even}\}}$ is 1 if k is even and 0 otherwise. It is this form of $b^{(k)}$ that was used in our computations.

A.2.2 Proof of Equations (12) and (13)

For a more compact notation, define $b^{(k)} \equiv b^{(k)}(\lambda_L(1-\theta))$ and $b \equiv b^{(0)}$. Now we show that, for $k = 1, 2, \ldots,$

$$
\begin{aligned}
&\frac{\Pi_L^{(k)}(\theta)(b-\theta)}{(1-\rho)} = \\
&\quad \frac{1}{(1-\rho)} \left(k \Pi_L^{(k-1)}(\theta) - \sum_{j=0}^{k-1} \binom{k}{j} (-\lambda_L)^{k-j} b^{(k-j)} \Pi_L^j(\theta) \right) \\
&\quad + (-\lambda_L)^{k-1} \left((2-k) b^{(k-1)} + (2 - \lambda_H - \lambda_L(1-\theta)) b^{(k)} \right)
\end{aligned}
\tag{40}
$$

Note that equation (13) follows from equation (40) by equation (10): i.e. by solving for $\Pi_L^{(k)}(\theta)$ in equation (40), letting $\theta = 0$ and dividing by $k!$. Similarly, equation (12) follows when θ is set to 0 in equation (6).

Using $c(\theta) = b(\theta)$ and rearranging terms in equation (6), we obtain

$$
\frac{\Pi_L(\theta)(b-\theta)}{1-\rho} = (1-\theta) b + \frac{\lambda_H}{\lambda_L} \cdot (1-b) b.
$$

Differentiating both sides with respect to θ yields

$$
\begin{aligned}
\frac{\Pi_L^{(1)}(\theta)(b-\theta)}{(1-\rho)} &= \frac{(1+\lambda_L b^{(1)})\Pi_L(\theta)}{1-\rho} - b - \lambda_L(1 - \theta + \frac{\lambda_H}{\lambda_L}) b^{(1)} + 2\lambda_H b b^{(1)} \\
&= \frac{(1+\lambda_L b^{(1)})\Pi_L(\theta)}{1-\rho} + b + (2 - \lambda_L(1-\theta) - \lambda_H) b^{(1)},
\end{aligned}
\tag{41}
$$

where the second line follows from the fact that $\lambda_H b b^{(1)} = b + b^{(1)}$, which can be checked using equation (38). This proves equation (40) for $k = 1$. Assume equation (40) holds for some $k \geq 1$. Define

$$
A(k) \equiv (-\lambda_L)^{k-1} \left((2-k) b^{(k-1)} + (2 - \lambda_H - \lambda_L(1-\theta)) b^{(k)} \right).
$$

Then equation (40) can be rewritten as

$$\begin{aligned}(1-\rho)A(k) &= \Pi_L^{(k)}(\theta)(b-\theta) - k\Pi_L^{(k-1)}(\theta) \\ &\quad + \sum_{j=0}^{k-1}\binom{k}{j}(-\lambda_L)^{k-j}b^{(k-j)}\Pi_L^{(j)}(\theta).\end{aligned}$$

Take the derivative with respect to θ of both sides. The derivative of the RHS is just the RHS with $k+1$ in place of k; the required steps are similar to those used to prove the formula for $b^{(k)}(\theta)$. For the LHS:

$$\begin{aligned}\frac{dA(k)}{d\theta} &= (-\lambda_L)^{k-1}[-\lambda_L(2-k)b^{(k)} + \lambda_L b^{(k)} - \lambda_L(2-\lambda_H-\lambda_L(1-\theta))b^{(k+1)}] \\ &= (-\lambda_L)^k[(2-k)b^{(k)} - b^{(k)} + (2-\lambda_H-\lambda_L(1-\theta))b^{(k+1)}] \\ &= A(k+1).\end{aligned}$$

Thus equation (40) holds for $k+1$ and hence, by induction, for all $k \geq 1$. Thus, equation (13) holds.

A.3 Proof of Proposition 1

The formula for $P[Q_H = 0]$ follows from equation (9) by setting $\theta = 0$. The formula for the Q_H probabilities, equation (11), is proved the same way as the formula for the Q_L probabilities — by proving by induction the following recursive formula for the generating function for $k \geq 1$:

$$\begin{aligned}&(1-\rho_H)(-\lambda_H)^{k-1}\{kU_H^{(k-1)} + \lambda_H(1-\theta)U_H^{(k)}\} = \\ &\quad -\Pi_H^{(k)}(\theta)(U_H-\theta) + k\Pi_H^{(k-1)}(\theta) - \sum_{j=1}^{k}\binom{k}{j}(-\lambda_H)^j U_H^{(j)}\Pi_H^{(k-j)}(\theta),\end{aligned}$$

where $U_H^{(k)} \equiv U_H^{(k)}(\lambda_H(1-\theta))$ and $U_H \equiv U_H^{(0)}$.

The case $k = 1$ follows by differentiating equation (9) with respect to θ. For the inductive step, one will find each side of the formula is self-differentiable; i.e. that the derivative of a side yields the same expression but with $k+1$ in place of k. The derivative of the right hand side uses steps similar to those used in the proof of the formula for $b^{(k)}(\theta)$.

References

[1] Chance, F. 1993. *Delphi: A C-Based Queueing Network Simulator.* Technical Report No. 1045, School of Operations Research and Industrial Engineering, Cornell University.

[2] Cohen, J.W. 1969. *The Single Server Queue.* North-Holland, Amsterdam.

[3] Federgruen, A. and P. Zipkin. 1986. An Inventory Model with Limited Production Capacity and Uncertain Demands. I. The Average-Cost Criterion. *Mathematics of Operations Research.* **11**, 193–207.

[4] Federgruen, A. and P. Zipkin. 1986. An Inventory Model with Limited Production Capacity and Uncertain Demands. II. The Discounted-Cost Criterion. *Mathematics of Operations Research.* **11**, 208–215.

[5] Heyman, D.P. 1968. Optimal Operating Policies for M/G/1 Queueing Systems. *Operations Research.* **16**, 362-382.

[6] Gavish, B. and S.C. Graves. 1980. A One-Product Production/Inventory Problem under Continuous Review Policy. *Operations Research.* **28**, 1228-1235.

[7] Güllü, A.R. and P.L. Jackson. 1993. *On the Continuous Time Capacitated Production/Inventory Problem with No Set Up Costs.* Technical Report No. 1054, School of Operations Research and Industrial Engineering, Cornell University.

[8] Jaiswal, N.K. 1968. *Priority Queues.* Academic Press, New York.

[9] Muckstadt, J.A. and R.O. Roundy. 1988. *Technical Report No. 806: Analysis of Multistage Production Systems.* School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY.

[10] Peterson, R. and E.A. Silver. 1979. *Decision Systems for Inventory Management and Production Planning.* Wiley, New York.

[11] Prabhu, N.U. 1965. *Queues and Inventories, a Study of their Basic Stochastic Processes.* Wiley, New York.

[12] Resnick, S. 1992. *Adventures in Stochastic Processes.* Birkenhäuser, Boston.

[13] Rudin, W. 1976. *Principles of Mathematical Analysis.* McGraw-Hill Book Company, New York.

[14] Schruben, L. 1991. *Sigma: A Graphical Simulation System.* Scientific Press, San Francisco.

[15] Sobel, M.J. 1969. Optimal Average-Cost Policy For a Queue with Start-Up and Shut-Down Costs. *Operations Research.* **17**, 145-162.

[16] Tayur, S.R. 1992. *Computing the Optimal Policy for Capacitated Inventory Models.* Technical Report, Carnegie Mellon University.

[17] Wagner, H.M. and T.M. Whitin. 1958. Dynamic Version of the Economic Lot Size Model. *Management Science.* **5**, 89–96.

[18] Williams, T. 1984. Special Products and Uncertainty in Production/Inventory Systems. *European Journal of Operations Research.* **15**, 46-54.

[19] Zipkin, P. 1986. Models for Design and Control of Stochastic, Multi-item Batch Production Systems. *Operations Research.* **34**, No. 1.

[20] Zheng, Y. and P. Zipkin. 1990. A Queueing Model to Analyze the Value of Centralized Inventory Information. *Operations Research.* **38**, No. 2.