ALLOCATING PRODUCTION CAPACITY AMONG MULTIPLE PRODUCTS

PAUL GLASSERMAN

Columbia Business School, New York, New York

(Received July 1993; accepted June 1994.)

We consider the problem of allocating production capacity among multiple items, assuming that a fixed proportion of overall capacity can be dedicated exclusively to the production of each item. Given a capacity allocation, production of each item follows a base-stock policy, i.e., each demand triggers a replenishment order to restore safety stocks to target levels. We present procedures for choosing base-stock levels and capacity allocations that are asymptotically optimal. Our objective is to minimize holding and backorder costs, or to minimize holding costs subject to a service-level constraint. *Asymptotic* optimality refers to large backorder penalties or stringent service-level constraints. Numerical results indicate that our rules perform very well even far from the asymptotic regime. A further approximation step results in allocation rules based on heavy-traffic limits; these, too, perform well.

We consider a manufacturer producing several items and keeping safety stocks of each item to supply variable external demands. An overall production capacity is to be allocated among the various items to minimize inventory costs; these costs may be holding and shortage costs, or holding costs to meet a target service level. We develop efficiently computable allocation rules that are asymptotically optimal as backorder penalties become very large or target service levels become very high.

The setting we consider is illustrated in Figure 1. The manufacturer receives orders for the items it produces; an order may be for a single item or for multiple units of multiple items. Production is *make-to-stock*, with each demand for an item triggering a replenishment order for that item.

This general model is sufficiently flexible to accommodate the following specific interpretations:

1. The items are distinct components assembled into a single product, or kit, at a single facility. In this case, an order is represented by a vector $(x_1, \ldots, x_d)$ in which x_i records the number of type-i components required for the assembly, and d is the total number of items. More generally, if the items are assembled into k products the demand vector takes k possible values; the probability of the jth value is the proportion of orders for the jth product, $j = 1, \ldots, k$.

2. The items are distinct products produced at a single facility and supplying distinct demands. In this case, the demand distribution factors into the product of its marginals, and there is no dependence across items.

3. The items all represent the same product manufactured at and distributed from different locations.

4. The items all represent the same product, but different inventories are kept for high-priority and low-priority customers.

Other variants are possible as well. Throughout, our analysis is motivated primarily by the first two cases above.

Overall capacity is measured by a maximum production rate r, and this rate is to be allocated among the various items. A premise of our analysis is that item i may be allocated a rate r_i, with $\Sigma_i\, r_i = r$, which is then always available for the production of that item and only of that item. This is a close representation of reality if the items are produced by distinct facilities; it is a somewhat cruder approximation of reality if, instead, there is just one facility. For in that case the full capacity would typically be devoted to a single item at a time and an allocation achieved through the time devoted to each item. We view our model as a sensible approximation at an aggregate level of the single-facility setting, in much the same way that the processor-sharing model of queueing theory is commonly taken as an approximation of a round-robin discipline. The allocated rates r_i, or more precisely the proportions r_i/r, may be viewed as surrogates for the proportion of time spent on each item over a base period, e.g., one week. This approximation may be too crude for job sequencing, but is reasonable for planning the overall effort to be dedicated to each item.

Taking our model a step further allows us to incorporate set-up times. Suppose, for example, that items are produced in a simple cyclic schedule, with the time slot dedicated to item i a fraction r_i/r of the time to complete a full cycle. Once the job sequence is fixed, so is the sequence of set-ups, and presumably so too is the total set-up time in a cycle. This time may be subtracted from the capacity r, and the allocation problem solved for the remaining *effective* capacity. This variant does not change our analysis so we do not pursue it further.

Our static allocation problem contrasts with the dynamic scheduling problems addressed in Wein (1992) and Zheng and Zipkin (1990) and the priority scheme in Carr et al. (1993). Our problem is closer to those solved by Kleinrock (1976, Chapter 5) and Wein (1989), but differs in one important respect: we cannot base our allocation on an

Subject classifications: Inventory/production approximation; asymptotically optimal allocations. Probability: stochastic model applications, production allocation. Programming: resource allocation.
Area of review: STOCHASTIC PROCESSES AND THEIR APPLICATIONS.

0030-364X/96/4405-0724 $01.25

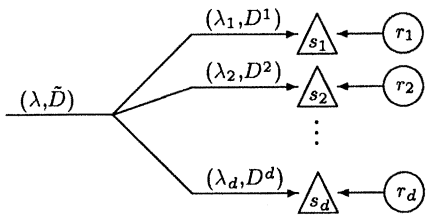

Figure 1. A facility producing multiple items. Orders arrive at rate λ and have the distribution of the vector $\tilde{D}$. Orders for item i arrive at rate λ_i and have the distribution of D^i. Item i is produced at rate r_i to restore inventory to the target level s_i.

explicit objective function (as they do) because no closed-form objective is available in our setting. To make the problem tractable, we consider a limiting regime in which backorder penalties become very large or service-level constraints become very stringent. Passing to the limit yields allocation rules that are asymptotically optimal for the original problem. Our use of asymptotics is similar in spirit to Anantharam (1989), though that work uses closed-form expressions for Jackson networks. Anantharam allocates buffers among queues to (asymptotically) maximize the time to overflow; one interpretation of our simplest allocation rule is that it asymptotically maximizes the time between stockouts.

We formulate our allocation rules and state our main asymptotic optimality results in Section 1, following the introduction of model details and necessary notation. In Section 2, we analyze a simplified problem that treats all items equally. The objective in this case is to allocate overall levels of safety stock and production capacity to minimize the frequency of stockouts; we give an asymptotically optimal solution. Building on this simplified model, in Section 3 we prove our main results—asymptotic optimality of allocation rules that take account of costs and service levels. In Section 4 we show that a further simplification of our allocation problem results in allocation rules based on heavy-traffic limits. Section 5 presents numerical results; these indicate that our asymptotically optimal allocations perform very well, even far from the asymptotic regime.

1. THE MODEL AND MAIN RESULTS

As suggested by Figure 1, we think of demands as arriving in a single stream, then splitting into demands for specific items. The times between arrivals of demands form an i.i.d. sequence having mean λ^{-1}. With d the number of items, a demand is characterized by a d-dimensional vector whose ith component is the quantity ordered of item i. The demand vectors form an i.i.d. sequence with d-dimensional joint distribution $F_{\tilde{D}}$. Demands are independent of interarrival times.

Since not all demands need include an order for item i, the epochs of arrivals of orders for item i form a subsequence of the epochs of arrivals of all demands; let $\{X_n^i, n \geq 0\}$ be the intervals between orders for item i. Our independence assumptions ensure that these interarrival times are i.i.d.; let λ_i^{-1} be their common mean. Denote by F_D^i the distribution of (genuine) demands for item i. More precisely, if $\tilde{D} = (\tilde{D}^1, \ldots, \tilde{D}^d)$ has distribution $F_{\tilde{D}}$, then

$$F_D^i(x) = P(\tilde{D}^i \leq x | \tilde{D}^i > 0), \quad x > 0,$$

and $F_D^i(0) = 0$.

To simplify notation, we assume throughout that inventories and demands are measured in units of work content. For example, if producing a type-1 item takes two hours at production rate 1 and producing a type-2 item takes three hours at that rate, then n units of inventory of each are recorded as $2n$ and $3n$, respectively, and similarly for units of demands. With this convention, let $m = \Sigma_i \mathsf{E}[\tilde{D}^i]$ be the total mean demand per order, $\tilde{D}$ having distribution $F_{\tilde{D}}$, and let $m_i = \mathsf{E}[D^i]$ be the mean demand per order for item i, D^i having distribution F_{D^i}. We always assume that

$$\lambda m < r, \tag{1}$$

meaning that there is sufficient capacity to meet demand. In allocating capacity r_i to item i, $i = 1, \ldots, d$, we only consider allocations satisfying

$$\lambda_i m_i < r_i, \quad i = 1, \ldots, d. \tag{2}$$

In other words, we are allocating the *excess* capacity

$$r - \sum_{i=1}^{d} \lambda_i m_i,$$

among the d items.

For any allocation $(r_1, \ldots, r_d)$, the operation of the system is determined by a vector $(s_1, \ldots, s_d)$ of *base-stock levels*. Item i is produced at rate r_i until its inventory reaches s_i, at which point production of the item ceases. Each demand for item i triggers an immediate replenishment order and thus re-initiates production, again at rate r_i. Demands not met from stock are backordered. For a single item in isolation, Federgruen and Zipkin (1986) establish the optimality of such a base-stock policy in a closely related periodic-review model; see Tayur (1993) for computational considerations. Akella and Kumar (1986) establish the optimality of an analogous policy in a model with deterministic demand and random breakdowns. Interestingly, the form of our asymptotically optimal base-stock level, given in (8), is superficially similar to the optimal inventory level of Akella and Kumar.

Denote by I_n^i the *net inventory* (on-hand minus backorders) of item i just prior to the arrival of the nth order for that item. Between orders, items of type i are produced at rate r_i until the target s_i is reached. Thus,

$$I_{n+1}^i = \min \{s_i, I_n^i - D_n^i + r_i X_n^i\}, \tag{3}$$

and this recursion completely specifies the evolution of the system under allocation $(r_1, \ldots, r_d)$. Let J_t^i be the net inventory of item i at time t; then J_t^i and I_n^i coincide at the epochs of arrival of orders for item i. Between orders, J_t^i increases linearly at rate r_i until it reaches s_i.

Under condition (2), the processes $\{(I_n^1, \ldots, I_n^d), n \geq 0\}$ and $\{(J_t^1, \ldots, J_t^d), t \geq 0\}$ admit unique stationary distributions; let $(I^1, \ldots, I^d)$ and $(J^1, \ldots, J^d)$ be random vectors with those distributions. Suppose inventory of item i is charged a holding cost at rate h_i and backorders of that item are penalized at rate p_i. Then the long-run average cost associated with item i at a base-stock level of s_i is

$$v_i(s_i) = h_i \mathsf{E}[(J^i)^+] + p_i \mathsf{E}[(J^i)^-], \tag{4}$$

and the total long-run average cost for the system is

$$v(s) = \sum_i v_i(s_i).$$

The long-run average proportion of orders for item i fully met from stock is

$$\alpha_i(s_i) = P(I^i > D^i), \tag{5}$$

where D^i has distribution F_{D^i} and is independent of I^i. Our objective is to choose base-stock levels $(s^1, \ldots, s^d)$ and a capacity allocation $(r_1, \ldots, r_d)$ to minimize the average cost or to minimize holding costs subject to a constraint on α_i.

These problems are intractable in the generality in which we have formulated them. We replace them with simpler problems and show that these surrogates provide asymptotically optimal solutions to the original problems. For the asymptotics, we impose some relatively minor assumptions on the interarrival and demand distributions. To simplify our limits, we require that, for each i, either D^i or X^i have a continuous distribution. (Without this condition, our limits hold through appropriate subsequences, as in the renewal theorem for arithmetic distributions.) Next, observe that any allocation satisfying (2) necessarily satisfies

$$r_i \leq \bar{r}_i \stackrel{\Delta}{=} r - \sum_{j \neq i} \lambda_j m_j;$$

we require that

$$P(D^i - \bar{r}_i X^i > 0) > 0. \tag{6}$$

This says that even at the maximal allocation of capacity to item i, there is some chance of a stockout of that item. Though not strictly necessary, this condition rules out certain trivial cases in which there is simply too much capacity for the allocation problem to be interesting. Finally, our main requirement is that for all r_i in the interval $(\lambda_i m_i, \bar{r}_i)$ there exist a $\theta_i > 0$ at which

$$1 < \phi_i(\theta_i) < \infty,$$

where

$$\phi_i(\theta) = \mathsf{E}[\exp\{\theta(D^i - r_i X^i)\}],$$

is the moment generating function of $D^i - r_i X^i$. This then implies the existence of exactly one $\gamma_i > 0$, for each r_i, at which

$$\phi_i(\gamma_i) = 1. \tag{7}$$

The functions $\gamma_i(r_i)$ are the key to our approach.

Consider the following resource allocation problem:

$$P_p\text{: minimize } \sum_i \frac{h_i}{\gamma_i(r_i)} \log\left(\frac{p_i + h_i}{h_i}\right),$$
$$\sum r_i = r,$$
$$r_i \geq \lambda_i m_i.$$

Let $\hat{r}$ solve problem P_p. For any $(r_1, \ldots, r_d)$, define

$$\tilde{s}_i = \frac{1}{\gamma_i(r_i)} \log\left(\frac{p_i + h_i}{h_i}\right). \tag{8}$$

Consider a sequence of models, indexed by n, with penalties $p_i^{(n)}$, $i = 1, \ldots, d$, all increasing to infinity as n increases. Let $v_*^{(n)}(r)$ be the minimum achievable cost under allocation $r = (r_1, \ldots, r_d)$ and let $\tilde{v}^{(n)}(r)$ be the cost achieved using the base-stock levels $(\tilde{s}_1, \ldots, \tilde{s}_n)$ defined by (8). The allocations provided by (8) and problem P_p are asymptotically optimal in the following sense:

Theorem 1. (i) *For any sequence $r^{(n)}$ of capacity allocations, $v_*^{(n)}(r^{(n)})/\tilde{v}^{(n)}(r^{(n)}) \to 1$.*

(ii) *If $\hat{r}^{(n)}$ solves P_p with backorder penalties $p^{(n)}$, and if $r^{(n)}$ is any other sequence of allocations, then*

$$\limsup_{n\to\infty} \frac{v_*^{(n)}(\hat{r}^{(n)})}{v_*^{(n)}(r^{(n)})} \leq 1.$$

Consider, next, the problem of minimizing holding costs subject to a service-level constraint. For each item i, we specify a maximum long-run stockout frequency δ_i, $0 < \delta_i < 1$; that is, we require $1 - \alpha_i(s_i) \leq \delta_i$. (Other standard measures of service are easily accommodated as well.) We introduce the following problem:

$$P_\delta\text{: minimize } \sum_i \frac{h_i}{\gamma_i(r_i)} \log\left(\frac{B_i}{\delta_i}\right)$$
$$\sum r_i = r$$
$$r_i \geq \lambda_i m_i,$$

for constants B_i to be specified in Section 2. Let

$$\tilde{s}_i = \frac{1}{\gamma_i(r_i)} \log\left(\frac{B_i}{\delta_i}\right). \tag{9}$$

Consider a sequence of models, indexed by n, with penalties $\delta_i^{(n)}$, $i = 1, \ldots, n$, all decreasing to zero as $n \to \infty$. Since service is constrained, set all backorder penalties to zero. Let $v_*^{(n)}(r)$ be the minimum achievable cost under allocation r, subject to the constraints $1 - \alpha_i(s_i) \leq \delta_i$, $i = 1, \ldots, d$, and let $\tilde{v}^{(n)}(r)$ be the cost achieved using base-stock levels $(\tilde{s}_1, \ldots, \tilde{s}_n)$ defined by (9). Then we have

Theorem 2. (i) *With $\tilde{s}_i^{(n)}$ as in (8), $1 - \alpha_i^{(n)}(\tilde{s}_i^{(n)}) \leq \delta_i^{(n)}$ for all n, and for any sequence $r^{(n)}$ of capacity allocations, $v_*^{(n)}(r^{(n)})/\tilde{v}^{(n)}(r^{(n)}) \to 1$.*

(ii) *If $\hat{r}^{(n)}$ solves P_δ and if $r^{(n)}$ is any other sequence of allocations, then*

$$\limsup_{n\to\infty} \frac{v_*^{(n)}(\hat{r}^{(n)})}{v_*^{(n)}(r^{(n)})} \leq 1 .$$

We prove Theorems 1 and 2 in Section 3. For these to be useful, the optimizations they entail—problems P_p and P_δ—must admit efficient solution procedures. In Section 2 we give a widely applicable sufficient condition for $\gamma_i^{-1}(\cdot)$ to be convex, and thus for P_p and P_δ to be separable, convex resource allocation problems. These are among the simplest nonlinear programming problems and efficient algorithms for their solution have been studied extensively; see, in particular, Luss and Gupta (1975), Zipkin (1980), and Sections 2.2–2.3 of Ibaraki and Katoh (1988). Thus, we have replaced the original allocation problems with tractable surrogates while preserving (asymptotic) optimality.

2. MINIMIZING THE STOCKOUT FREQUENCY

Before justifying the rules set forth in the previous section, we treat a simplified problem which is of independent interest and, more importantly, lends insight into the approach that follows. We treat the items symmetrically—omitting holding costs, penalties, and service levels from the discussion—and allocate capacity to minimize the overall frequency of stockouts, in an asymptotic sense.

Our analysis simplifies if instead of the net inventory for each item we record the *shortfall*, which is the difference between the base-stock level and the net inventory. Just prior to the arrival of the nth order for item i, the shortfall in that item is

$$Y_n^i = s_i - I_n^i .$$

It follows from (3) that this quantity satisfies the Lindley equation

$$Y_{n+1}^i = \max\{0,\ Y_n^i + D_n^i - r_i X_n^i\}, \quad i = 1, \ldots, d . \tag{10}$$

Thus, $\{Y_n^i\}$ coincides with the waiting-time process in a queue with service times $\{D_n^i\}$ and interarrival times $\{r_i X_n^i\}$. Under condition (2), the shortfall process has a stationary distribution; let $(Y^1, \ldots, Y^d)$ have that distribution. Similarly, let $V_t^i = s^i - J_t^i$ be the continuous-time shortfall in item i and let $(V^1, \ldots, V^d)$ have the corresponding stationary distribution. The distribution of V^i is that of the steady-state workload (virtual waiting time) in the associated queue.

We can express $\alpha_i(s_i)$, the steady-state probability that an order for item i is met from stock, as

$$\alpha_i(s_i) = P(Y^i + D^i \leq s_i) , \tag{11}$$

where D^i has distribution F_D^i and is independent of Y^i. It follows from general results on reflected random walks and queues that, for each item i, there is a constant C_i such that

$$1 - \alpha_i(s_i) \sim C_i e^{-\gamma_i s_i} , \tag{12}$$

where the symbol $\sim$ means that the ratio of the two sides converges to unity as s_i increases to infinity, and γ_i is as defined in (7). See, e.g., Asmussen (1987, p. 269), Feller (1971, §XII.6), or Siegmund (1985, p. 175) for a corresponding result for $P(Y^i > s^i)$; the extension to (12) merely modifies the constant C_i. Furthermore, it follows from Kingman (1970) and Ross (1974) that there are positive constants A_i, B_i such that

$$A_i e^{-\gamma_i s_i} \leq 1 - \alpha_i(s_i) \leq B_i e^{-\gamma_i s_i} , \tag{13}$$

for all $s_i > 0$. In fact, we can (and do) choose A_i, B_i independent of r_i; for example, we may take

$$A_i = \left(\sup_{x\geq 0} \mathsf{E}[\exp\{\gamma_i(D^i - x)\} | D^i > x] \right)^{-1} ,$$

and $B_i = \phi_{D^i}(\gamma_i(\bar{r}_i))$, with ϕ_{D^i} the moment generating function of D^i. This A_i is shown in Ross to provide a lower bound for $\exp(\gamma_i s^i) P(Y^i > s^i)$. Kingman shows that $P(Y^i > s_i) \leq \exp(-\gamma_i s_i)$; multiplying by our B_i makes the bound valid for $1 - \alpha_i(s_i)$. The precise values of the constants in (12) and (13) are less important than the existence of *some* constants making these expressions valid. We use properties (12) and (13) to identify an asymptotically optimal capacity allocation.

Our approach proceeds in two steps. First, for any fixed capacity allocation we identify an asymptotically optimal allocation of safety stock to the d items. Then, we pick the capacity allocation that optimizes this asymptotic optimum. We therefore begin by fixing a capacity allocation $(r_1, \ldots, r_d)$ and considering a sequence of base-stock vectors $(s_1, \ldots, s_d)$ with $\sum_i s_i = s$ and s increasing to infinity. Let $\alpha(s)$ be the steady-state probability that an arriving order (possibly for multiple items) is fully met from stock. This probability behaves as follows:

Proposition 1. *For each allocation $r = (r_1, \ldots, r_d)$, there exist positive constants A, B such that*

$$A \exp(-\min_i \{\gamma_i s_i\}) \leq 1 - \alpha(s)$$
$$\leq B \exp(-\min_i \{\gamma_i s_i\}) , \tag{14}$$

for all s. If there exist q_i, $i = 1, \ldots, d$, for which

$$\lim_{s\to\infty} s_i/s = q_i, \quad i = 1, \ldots, d ,$$

then

$$\lim_{s\to\infty} -s^{-1} \log(1 - \alpha(s)) = \min_i \{\gamma_i q_i\} ;$$

i.e., $\min_i\{\gamma_i q_i\}$ is the exponential rate at which the stockout probability vanishes as the overall level of safety stock increases.

Proof. The stockout probability $1 - \alpha(s)$ satisfies

$$1 - \alpha(s) = P(\text{some item } i \text{ with } \tilde{D}^i > 0 \text{ stocks out})$$
$$\leq \sum_i P(\tilde{D}^i > 0 \text{ and item } i \text{ stocks out})$$
$$= \sum_i P(\tilde{D}^i > 0) P(Y^i + \tilde{D}^i > s_i | \tilde{D}^i > 0)$$
$$= \sum_i P(\tilde{D}^i > 0) P(Y^i + D^i > s_i)$$
$$\leq \left(\sum_i P(\tilde{D}^i > 0) B_i \right) \exp\left(-\min_i \{\gamma_i s_i\}\right) ,$$

so we may take $B = \Sigma_i P(\tilde{D}^i > 0)B_i$. For the lower bound, fix an arbitrary item i and observe that

$$\begin{aligned} 1 - \alpha(s) &\geq P(\tilde{D}^i > 0 \text{ and item } i \text{ stocks out}) \\ &= P(\tilde{D}^i > 0)(1 - \alpha_i(s_i)) \\ &\geq P(\tilde{D}^i > 0)A_i \exp(-\gamma_i s_i) . \end{aligned}$$

Setting $A = \min_i P(\tilde{D}^i > 0)A_i$, we conclude that

$$1 - \alpha(s) \geq A \exp(-\gamma_i s_i), \quad i = 1, \ldots, d ;$$

taking the maximum over i concludes the proof of the lower bound. The second part of the proposition follows by taking logarithims in the upper and lower bounds and letting $s \to \infty$. □

This result suggests the following rule for allocating an overall safety stock level s among the d items: choose the proportion q_i of stock allocated to item i to maximize the rate $\min_i \{\gamma_i q_i\}$ at which the stockout probability vanishes. Choosing q_i subject to $\Sigma_i q_i = 1$ is equivalent to choosing s_i subject to $\Sigma_i s_i = s$, so we arrive at the following minimax (maximin) allocation problem:

$$\begin{aligned} P_s\text{: maximize } & \min \{\gamma_i s_i\} , \\ & \sum s_i = s , \\ & s_i \geq 0 . \end{aligned}$$

The optimal solution evidently sets $\gamma_i s_i$ constant across items; i.e.,

$$s_i = \left(\sum_j \gamma_j^{-1} \right)^{-1} s \gamma_i^{-1} , \tag{15}$$

and results in the objective-function value of s times

$$\left(\sum_j \gamma_j^{-1} \right)^{-1} . \tag{16}$$

Thus, for a given $(r_1, \ldots, r_d)$, (16) gives the maximal exponential rate of decrease of the stockout probability as the overall safety stock level increases.

Having identified an asymptotically optimal allocation of safety stocks, we now turn to the problem of choosing $(r_1, \ldots, r_d)$. The rate identified in (16) suggests the following rule: allocate capacity to maximize the maximum rate of decrease of the stockout probability. Maximizing (16) is equivalent to minimizing its reciprocal; thus, we arrive at the following resource allocation problem:

$$\begin{aligned} P_\alpha\text{: minimize } & \sum_i \gamma_i^{-1}(r_i) , \\ & \sum r_i = r , \\ & r_i \geq \lambda_i m_i . \end{aligned}$$

Each γ_i^{-1} is continuous and increases to infinity as r_i approaches $\gamma_i m_i$, so problem P_α has a solution. Any solution to this problem has the following property:

Theorem 3. *Let r^* solve problem P_α and let $1 - \alpha_{r^*}(s)$ be the stockout probability under capacity allocation r^* and stock allocation (15). Then $1 - \alpha_{r^*}(s)$ is the asymptotically smallest stockout probability; indeed, if r fails to solve problem P_α and if $1 - \alpha_r(s)$ denotes the stockout probability under capacity allocation r and any stock allocation, then $(1 - \alpha_{r^*}(s))/(1 - \alpha_r(s)) \to 0$ exponentially fast.*

Proof. For any allocation r, let $\xi(r)$ be the corresponding value of the expression in (16). If r fails to solve problem P_α, then $\xi(r) \leq \xi(r^*)$. Moreover, from Proposition 1, we know that

$$1 - \alpha_{r^*}(s) \leq B \exp(-\xi(r^*)s) ,$$

whereas

$$1 - \alpha_r(s) \geq A \exp(-\min_i \{\gamma_i s_i\}) \geq A \exp(-\xi(r)s) ,$$

so the ratio is $O(\exp[-(\xi(r^*) - \xi(r))s])$ as $s \to \infty$. □

Remarks. (i) Minimizing the asymptotic stockout probability is roughly equivalent to maximizing the asymptotic time between stockouts, and this is in the same spirit as Anantharam's buffer allocation rule for Jackson networks: allocate buffers to maximize the time to overflow. A key step in his approach is identifying the exponential rate at which the overflow probability vanishes, based on an analysis of product-form distributions.

(ii) These allocation criteria are also counterparts of *Kelly gambling* criteria in which one maximizes the exponential growth rate of one's fortune; see, e.g., Breiman (1961).

(iii) While we have detailed only the stockout criterion, the same rule applies to the *fill rate* (proportion of demands met from stock) and the expected backorders. These measures of service have the same exponential decrease as the stockout probability; only the constants outside the exponential term are different (see Glasserman 1993).

For problem P_α and our other allocation rules to be of practical value, they must admit efficient solution procedures; as discussed in Section 1, they do if every γ_i^{-1} is convex. We now give a sufficient condition for convexity. Define

$$\psi_{X^i}(\theta) = \log \mathsf{E}[e^{\theta X^i}] , \tag{17}$$

and let $\psi_{X^i}^{\leftarrow}$ be the inverse mapping. The following result is proved in an appendix:

Proposition 2. *Suppose the nth derivative of $\psi_{X^i}^{\leftarrow}$ is positive for odd n and negative for even n. Then the mapping $r_i \mapsto \gamma_i^{-1}(r_i)$ is convex and consequently problem P_α is a separable, convex resource allocation problem.*

We have not been able to identify a distribution for which the logarithmic moment generating function (also called the cumulant generating function) defined in (17) fails to satisfy the requirement in Proposition 2. As a simple illustration, consider exponentially distributed interarrival times, for which

$$\psi(\theta) = \log\left(\frac{\lambda}{\lambda - \theta}\right) ,$$

and therefore

$$\psi^{\leftarrow}(u) = \lambda(1 - e^{-u}),$$

which does indeed have derivatives of alternating signs. The same is easily verified for the normal, Poisson, gamma (including Erlang), and inverse Gaussian distributions. Moreover, ψ_i is always increasing and convex, so $\psi_i^{\leftarrow}$ is always increasing and concave, and the requirement in Proposition 2 is thus automatic for $n = 1, 2$.

Returning to the allocation problem P_α, some insight is provided by the case of exponentially distributed interarrival times and order sizes. Suppose, then, that orders for item i are exponentially distributed with mean μ_i^{-1}. In this case, γ_i is defined by the equation

$$\left(\frac{\lambda_i r_i^{-1}}{\lambda_i r_i^{-1} + \gamma_i}\right)\left(\frac{\mu_i}{\mu_i - \gamma_i}\right) = 1,$$

resulting in

$$\gamma_i = \mu_i - \lambda_i r_i^{-1}. \tag{18}$$

It is not hard to see that for this γ_i, problem P_α is solved by setting

$$r_i = \frac{\lambda_i + \nu\sqrt{\lambda_i}}{\mu_i}, \tag{19}$$

with ν chosen to make the r_i's sum to r; i.e.,

$$\nu = \frac{r - \sum_i \lambda_i \mu_i^{-1}}{\sum_i \sqrt{\lambda_i}\mu_i^{-1}}.$$

If all μ_i are equal to one, then this rule sets the capacity for item i equal to the mean demand per unit time for item i plus a fixed number of standard deviations. Except for minor differences in the models, this coincides with the capacity assignment in Kleinrock based on minimization of an explicit objective function available in the exponential case. See Reiman (1990) and Wein (1989) for related results based on heavy-traffic objective functions.

The solution in (19) should be contrasted with the more straightforward proportional allocation that makes $\lambda_i/(\mu_i r_i)$ constant across i. Rather than set r_i proportional to λ_i/μ_i, the rule derived above sets the *slack* $r_i - (\lambda_i/\mu_i)$ proportional to $\sqrt{\lambda_i}/\mu_i$. Numerical results in Section 5 indicate that proportional allocations perform far less well than our asymptotically optimal rules.

3. COSTS AND CONSTRAINTS

We now adapt the basic allocation strategy developed in the previous section to account for holding costs, backorder penalties and service-level constraints.

3.1. Minimizing Costs

The basic allocation rule in Section 2 builds on the observation that the stockout probability for the ith item vanishes at the exponential rate γ_i as the base-stock level s_i increases. The same principle leads to an allocation rule in the presence of holding costs and backorder penalties. To make this connection, we need some further properties of individual inventories.

As in Section 1, let $(V^1, \ldots, V^d)$ have the stationary distribution of the continuous-time shortfall process. It follows from (2), (7), and general results relating V and Y (see, e.g., p. 189 of Asmussen) that there is a constant C_i such that

$$P(V^i > x) \sim C_i e^{-\gamma_i x}, \quad \text{as } x \to \infty. \tag{20}$$

The constant C_i featured here is not in general the same as the one in (12); however, since the precise values of the constants play no role in our analysis we use the same symbol for both. As before, there are constants A_i, B_i, not depending on r_i, such that

$$A_i e^{-\gamma_i x} \leq P(V^i > x) \leq B_i e^{-\gamma_i x}, \quad \text{for all } x > 0. \tag{21}$$

We may choose these constants and the ones in (13) to coincide by taking the smaller of the A_i's and the larger of the B_i's in each case, but again the precise values are not important.

Now suppose, as in Section 1, that inventories of item i are charged a holding cost at rate h_i and that backorders are penalized at rate p_i. Then the long-run average cost associated with item i, viewed as a function of the base-stock level s_i, can be expressed as

$$v_i(s_i) = h_i \mathsf{E}[(s_i - V^i)^+] + p_i \mathsf{E}[(V^i - s_i)^+]. \tag{22}$$

Differentiating and setting the derivative equal to zero shows that the average cost is minimized at the point s_{*i} satisfying

$$P(V^i > s_{*i}) = \frac{p_i}{h_i + p_i}. \tag{23}$$

Properties (20) and (21) imply the following characterization of s_{*i}:

Lemma 1. *With $B_i \geq 1$, for all $p_i > 0$ we have*

$$\gamma_i^{-1} \log\left(\frac{A_i(p_i + h_i)}{h_i}\right) \leq s_{*i} \leq \gamma_i^{-1} \log\left(\frac{B_i(p_i + h_i)}{h_i}\right);$$

and as $p_i \to \infty$,

$$\left| \gamma_i^{-1} \log\left(\frac{C_i(p_i + h_i)}{h_i}\right) - s_{*i} \right| \to 0.$$

A similar result is established in Glasserman for a periodic-review model, so we omit the proof. We need a related property of the cost function for large backorder penalties. Rewriting the ith cost function as

$$v_i(s_i) = h_i(s - \mathsf{E}[V^i]) + (h_i + p_i)\mathsf{E}[(V^i - s_i)^+], \tag{24}$$

shows that the behavior of the expected backlog $\mathsf{E}[(V^i - s_i)^+]$ is critical to the behavior of the expected cost. From (20) and (21), we obtain

Lemma 2. *For all $s_i > 0$,*

$$A_i \gamma_i^{-1} e^{-\gamma_i s_i} \leq \mathsf{E}[(V^i - s_i)^+] \leq B_i \gamma_i^{-1} e^{-\gamma_i s_i};$$

moreover,

$$\mathsf{E}[(V^i - s_i)^+] \sim C_i \gamma_i^{-1} e^{-\gamma_i s_i},$$

as $s_i \to \infty$.

This, too, parallels a result for the periodic-review model in Glasserman so we again omit the proof. We use the preceding lemmas to develop an allocation rule for the capacity r. As in the basic allocation problem of Section 2, we develop the rule in two steps: first, we identify (asymptotically) optimal base-stock levels for each allocation $(r_1, \ldots, r_d)$; then, we choose the allocation to optimize the asymptotic optimum. For any choice of s, we choose the capacity allocation that minimizes the growth of

$$H(s) \stackrel{\Delta}{=} \sum_i h_i s_i,$$

as the p_i's become large. This choice of objective is supported by our next lemma. As in Section 1, consider a sequence of problems indexed by n with $p_i^{(n)} \to \infty$ as $n \to \infty$. Let $v^{(n)}$ be the corresponding cost functions, let $s_*^{(n)}$ minimize $v^{(n)}$, and let $\tilde{s}^{(n)}$ be as defined in (8) with p_i replaced by $p_i^{(n)}$.

Lemma 3. *For any sequence* $r^{(n)}$ *of allocations,* $v^{(n)}(s_*^{(n)}) \sim H(s_*^{(n)})$ *and* $v^{(n)}(\tilde{s}^{(n)}) \sim H(\tilde{s}^{(n)})$.

Proof. To lighten notation, we omit the superscript n. It suffices to show that $v_i(s_{*i}) \sim h_i s_{*i}$ and $v_i(\tilde{s}_i) \sim h_i \tilde{s}_i$ for each i. From Lemma 2 and (24), we get

$$\begin{aligned} v_i(\tilde{s}_i) &\leq h_i(\tilde{s}_i - \mathsf{E}[V^i]) + (h_i + p_i)\gamma^{-1} B_i \exp[-\gamma_i \tilde{s}_i] \\ &= h_i(\tilde{s}_i - \mathsf{E}[V^i] + B_i \gamma_i^{-1}). \end{aligned}$$

Similarly, we have the lower bound

$$v_i(\tilde{s}_i) \geq h_i(\tilde{s}_i - \mathsf{E}[V^i]).$$

Dividing these bounds by $h_i \tilde{s}_i$ and passing to the limit proves the result for $v(\tilde{s})$.

For the optimal base-stock levels, we use the bounds in Lemma 1 to get

$$\begin{aligned} v_i(s_{*i}) &\leq h_i(s_{*i} - \mathsf{E}[V^i]) \\ &\quad + (h_i + p_i)\gamma^{-1} B_i \exp[-\gamma_i s_{*i}] \\ &\leq h_i(s_{*i} - \mathsf{E}[V^i] + B_i(A_i \gamma_i)^{-1}), \end{aligned}$$

and a corresponding lower bound. The result then follows as in the previous case. □

For optimal base-stock levels, it follows from the second part of Lemma 1 that

$$\left| H(s_*) - \sum_i \frac{h_i}{\gamma_i} \log\left(\frac{C_i(p_i + h_i)}{h_i}\right) \right| \to 0,$$

and this, together with Lemma 3, suggests that we choose $(r_1, \ldots, r_d)$ to minimize

$$\sum_i \frac{h_i}{\gamma_i} \log\left(\frac{C_i(p_i + h_i)}{h_i}\right). \tag{25}$$

In general, C_i is difficult to evaluate as a function of r_i, making (25) difficult to minimize. In addition, it is generally not possible to determine if this function is convex. We circumvent these difficulties by writing

$$\begin{aligned} \frac{h_i}{\gamma_i} \log\left(\frac{C_i(p_i + h_i)}{h_i}\right) &= \frac{h_i}{\gamma_i} \log(C_i) \\ &\quad + \frac{h_i}{\gamma_i} \log\left(\frac{p_i + h_i}{h_i}\right), \end{aligned}$$

and noting that only the second term changes with p_i. Thus, we arrive at the asymptotic objective function

$$\sum_i \frac{h_i}{\gamma_i} \log\left(\frac{p_i + h_i}{h_i}\right),$$

appearing in problem P_p. Moreover, under the condition in Proposition 2, problem P_p is a separable, convex resource allocation problem because only γ_i depends on r_i and we established in the proposition that γ_i^{-1} is convex.

We asserted in Theorem 1 that a solution to problem P_p gives an asymptotically optimal capacity allocation; we now justify this assertion.

Proof of Theorem 1. For part (i), consider an arbitrary sequence $r^{(n)}$ of allocations and let $v^{(n)}$ be the corresponding cost functions. Dropping the superscript n, we find from Lemma 3 that

$$\lim_{n\to\infty} \frac{v(r, s_*)}{v(r, \tilde{s})} = \lim_{n\to\infty} \frac{H(s_*)}{H(\tilde{s})} = 1, \tag{26}$$

the second equality following from

$$\lim_{n\to\infty} s_{*i}/\tilde{s}_i = \lim_{n\to\infty} \frac{\log((p_i + h_i)/h_i) + \log(C_i)}{\log((p_i + h_i)/h_i)} = 1.$$

For part (ii), let $\hat{r}^{(n)}$ solve problem P_p and let $r^{(n)}$ be any other sequence of allocations. Then

$$\limsup_{n\to\infty} \frac{v(\hat{r}, \tilde{s})}{v(r, s_*)} = \limsup_{n\to\infty} \frac{v(\hat{r}, \tilde{s})}{v(r, \tilde{s})} = \limsup_{n\to\infty} \frac{H(\hat{r}, \tilde{s})}{H(r, \tilde{s})} \leq 1,$$

the first equality following from (26), the second from Lemma 3, and the inequality from the fact that, by definition, $\hat{r}$ minimizes $H(\cdot, \tilde{s})$. □

In Section 5 we report numerical results based on this allocation rule showing that its performance in the original cost minimization problem is excellent.

3.2. Meeting a Service-level Constraint

We now carry out an analysis similar to that of Section 3.1 to develop a capacity allocation rule subject to service-level constraints. For each item i, let α_i be as in (11); then $1 - \alpha_i(s_i)$ is the long-run average proportion of orders for item i that cannot be fully met from stock. The service-level constraint we consider sets

$$1 - \alpha_i(s_i) \leq \delta_i, \quad i = 1, \ldots, d, \tag{27}$$

with each $0 < \delta_i < 1$ for each i. Other measures of service (in particular the fill rate) can be handled using similar

techniques, but we work with the stockout frequency because it is the simplest case.

Whereas in Section 3.1 s_{*i} denoted the cost-minimizing base-stock level for item i, we now take s_{*i} to be the smallest s_i satisfying (27). Much as in Lemma 1, we have

Lemma 4. *With $B_i \geq 1$, for all $0 < \delta_i < 1$,*

$$\gamma_i^{-1} \log (A_i/\delta_i) \leq s_{*i} \leq \gamma_i^{-1} \log (B_i/\delta_i) ,$$

and as $\delta_i \to 0$,

$$|\gamma_i^{-1} \log (C_i/\delta_i) - s_{*i}| \to 0 ,$$

where A_i, B_i, C_i are as in (12) and (13).

With the stockout probabilities constrained, we set the backorder penalties to zero, making the expected cost for item i equal to

$$\begin{aligned} v_i(s_i) &= h_i \mathsf{E}[(s_i - V^i)]^+ \\ &= h_i(s_i - \mathsf{E}[V^i] + \mathsf{E}[(V^i - s_i)^+]) . \end{aligned}$$

The properties of $\mathsf{E}[(V^i - s_i)^+]$ established in Lemma 2 are thus relevant here as well. If we set

$$\tilde{s}_i = \gamma^{-1} \log (B_i/\delta_i) ,$$

then, via Lemma 2, the conclusions in Lemma 3 of Section 3.1 continue to hold with s_* and $\tilde{s}$ as defined in this section. The steps used in Section 3.1 lead us to approximate the true holding cost by

$$H(\tilde{s}) = \sum_i h_i \tilde{s}_i = \sum_i h_i \gamma_i^{-1} \log (B_i/\delta_i) , \tag{28}$$

and to take minimization of (28) as the criterion for allocating capacity. This results precisely in problem P_δ introduced in Section 1. The proof of the asymptotic optimality claimed in Theorem 2 proceeds along the same lines as the proof of Theorem 1. The fact that the service level constraints are met by $\tilde{s}$ for all δ follows from the bound in Lemma 4:

$$1 - \alpha_i(\tilde{s}_i) \leq B_i \exp (-\gamma_i \tilde{s}_i) = \delta_i ,$$

by the definition of $\tilde{s}_i$.

4. ALLOCATION IN HEAVY TRAFFIC

A further simplification of the stock and capacity allocation problems is possible at high utilizations. Expanding the moment generating function ϕ_i appearing in (7) in a Taylor series about the origin yields

$$\phi_i(\theta) = 1 - \mu_i \theta + \tfrac{1}{2} \sigma_i^2 \theta^2 + o(\theta^2) ,$$

where

$$\mu_i = -\mathsf{E}[D^i - r_i X^i] = r_i \lambda_i^{-1} - m_i$$

and $\quad \sigma_i^2 = \mathrm{Var}\,[D^i - r_i X^i] .$

If $\lambda_i m_i / r_i$ is close to one, then μ_i is close to zero, so γ_i, the positive solution to $\phi_i(\gamma_i) = 1$, must be close to zero, indicating that

$$-\mu_i \gamma_i + \tfrac{1}{2} \sigma_i^2 \gamma_i^2 \approx 0 ;$$

that is,

$$\gamma_i \approx \eta_i \stackrel{\Delta}{=} 2\mu_i/\sigma_i^2 .$$

Thus, we have a two-moment approximation to γ_i that becomes increasingly accurate as the utilization increases to one. This suggests replacing γ_i with η_i in our allocation rules. A simple calculation shows that η_i^{-1} is always convex in r_i, so the resulting minimization problem is separable and convex.

Further justification for replacing γ_i with η_i follows from heavy-traffic limits for queues, as in, e.g., §VIII of Asmussen. For $\lambda_i m_i / r_i$ close to one, the distribution of V^i is close to the exponential distribution with mean $1/\eta_i$. If V^i had exactly this distribution, then the value of s_i minimizing v_i would be

$$s_i = \eta_i^{-1} \log \left(\frac{p_i + h_i}{h_i} \right) . \tag{29}$$

Moreover, the optimal capacity allocation would be obtained by minimizing

$$\sum_i \frac{h_i}{\eta_i(r_i)} \log \left(\frac{p_i + h_i}{h_i} \right) . \tag{30}$$

This is the objective function appearing in problem P_p with γ_i replaced by η_i. Expressions similar to (29) form part of the analysis in Wein (1992) of a dynamic scheduling problem. Wein (1989) uses a heavy-traffic approximation in allocating service rates in a network of queues to minimize the average number of jobs in the network. His objective function is a queueing-network counterpart of $\Sigma_i \eta_i^{-1}$.

While it is tempting to conjecture that allocations based on the heavy-traffic objective (30) are in some sense asymptotically optimal, it is unclear how such a result should even be formulated. The heavy-traffic asymptotics are fundamentally different from those of Sections 2 and 3 because *the conditions required for the limiting regime depend on the choice of allocation*. More specifically, the heavy-traffic limit requires $\lambda_i m_i \approx r_i$, a condition depending on r_i, whereas our previous limits required p_i large or δ_i small, regardless of the capacity allocation. Even if the overall utilization $\lambda m/r$ is close to 1, it is certainly possible that under an optimal allocation some $\lambda_i m_i / r_i$ would be much less than 1, making a heavy-traffic approximation questionable. Similar comments apply to the problem treated in Wein (1989). In spite of this theoretical shortcoming, numerical results in Section 5 indicate that replacing γ_i with η_i brings some simplification without much deterioration in performance.

It is less clear how to use heavy-traffic limits to minimize costs subject to service-level constraints, if these constraints are firm. In particular, if $\eta_i > \gamma_i$, then no base-stock level of the form $s_i = \eta_i^{-1} \log (b_i/\delta_i)$, with b_i a constant, will ensure a stockout frequency less than δ_i, as $\delta_i \to 0$.

5. EVALUATION OF THE ALLOCATION RULES

In order to compare our (asymptotically optimal) allocation rules with optimal allocations, we restrict attention to a tractable case: demands for item i arrive in a Poisson stream and order-sizes for item i are exponentially distributed. In this case, through the correspondence with workloads in queues, we know that the distribution of V^i is given by

$$P(V^i \leq x) = 1 - \rho_i \exp(-\gamma_i x), \tag{31}$$

with $\rho_i = \lambda_i m_i / r_i$ and $\gamma_i = m_i^{-1} - \lambda_i r_i^{-1}$; see, e.g., Prabhu (1980, p. 33). It follows that v_i is minimized at

$$s_i = \gamma_i^{-1} \log\left(\frac{\rho_i(p_i + h_i)}{h_i}\right), \tag{32}$$

resulting in the average cost

$$v_i(s_i) = \frac{h_i}{\gamma_i}\left[\log\left(\frac{\rho_i(p_i + h_i)}{h_i}\right) + (1 - \rho_i)\right].$$

The allocation $(r_1, \ldots, r_d)$ minimizing $\Sigma_i\, v_i(s_i)$ has the square-root form

$$\lambda_i m_i + \frac{m_i\sqrt{\lambda_i k_i}}{\sum_j m_j\sqrt{\lambda_j k_j}}\left(r - \sum_j \lambda_j m_j\right),$$

where

$$k_i = h_i\left[\log\left(\frac{\rho_i(p_i + h_i)}{h_i}\right) + (1 - \rho_i)\right].$$

The solution to our surrogate problem P_p has the same form but with k_i replaced by

$$\hat{k}_i = h_i \log\left(\frac{p_i + h_i}{h_i}\right).$$

Notice that $\hat{k}_i/k_i \to 1$ as either $\rho_i \to 1$ or $p_i \to \infty$.

We evaluated the performance of several allocation rules for a variety of cost structures and utilizations. The tractability of the exponential case allows us to compare our approximations with optimal costs. In addition to the asymptotic optimum derived in Section 3, we test the heavy-traffic approximation. In the exponential case, we have

$$\eta_i(r_i) = 2\left(\frac{r_i\lambda_i^{-1} - m_i}{r_i^2\lambda_i^2 + m_i^2}\right).$$

We also evaluated a proportional allocation in which the r_i's are selected to make all $\lambda_i m_i / r_i$ equal. For the asymptotic optimum and the heavy-traffic approximation we considered two cases: one using the approximately optimal base-stock levels (8) and (9), respectively, and one using the optimal levels (32). The first case tests the combination of the stock and capacity allocation rules, the second case tests the capacity allocation only. In total, we compared six allocation rules, given the following labels in the tables of results:

- Optimal: optimal r_i and s_i;
- Asy-Opt: asymptotically optimal r_i with optimal s_i;
- Asy-Asy: asymptotically optimal r_i and s_i;
- Hvy-Opt: heavy-traffic r_i with optimal s_i;
- Hvy-Hvy: heavy-traffic r_i and s_i;
- Proport: proportional r_i with optimal s_i.

All numerical results are based on three items, with corresponding arrival rates 0.8, 0.15, and 0.05. These values are representative of A-B-C classifications of products, A-products accounting for 80% of demand, B- and C-products for 15% and 5%, respectively; see, e.g., Carr et al. With this interpretation, our items become *groups* of products, and the problem becomes allocating stock and capacity among the groups. In all cases, we take the mean order-size per demand to be 1 and the holding-cost rate for each item to be 1 as well. We expect the qualitative effect of varying these parameters to be captured by the range of backorder penalties and utilizations we consider. Our experiments are divided into two cases:

Symmetric costs. In these experiments, backorder penalties are equal for all three products. With $\rho \equiv (\Sigma_i \lambda_i)/r = 1/r$, we consider the overall utilization levels $\rho = 0.5$, 0.7, and 0.9. Within each level, we take the (common) backorder penalty to be $p = 1$, 2, 5, 10, or 100.

Asymmetric costs. In these experiments we once again take ρ to be 0.5, 0.7, or 0.9. At each utilization level we consider six values of the penalty vector (p_1, p_2, p_3). The specific values appear along with the numerical results. In the first four cases, penalties either increase or decrease with the arrival rate; the last two cases are nonmonotone.

The results appear in Tables I and II. Overall, they show excellent performance for our allocation rules. Since the proportional allocation uses optimal base-stock levels, it should only be compared with Asy-Opt and Hvy-Opt. With this understanding, we summarize the results as follows:

- In all but four cases, Asy-Opt is less than 1% above Optimal; in all but one case it is less than 3% higher; in more than half the experiments, the two are virtually indistinguishable.
- Asy-Opt consistently outperforms Hvy-Opt and Asy-Asy consistently outperforms Hvy-Hvy.
- The Asy allocations depend less on the use of optimal base-stock levels than the Hvy allocations: in all but one case, the deterioration in performance in passing from Asy-Opt to Asy-Asy is less than that in passing from Hvy-Opt to Hvy-Hvy.
- The Asy allocations are less sensitive to conditions needed for the underlying approximation (large p) than the Hvy allocations (large ρ). This is best seen by comparing their performance (versus the optimum) at $\rho = 0.5, p = 100$, and $\rho = 0.9, p = 1$ in rows five and eleven of Table I.
- Proportional allocations are not competitive.

Table I
Symmetric Costs

ρ	p	Optimal	Asy-Opt	Asy-Asy	Hvy-Opt	Hvy-Hvy	Proport
0.5	1	1.75	1.92	3.65	2.57	4.61	3.00
	2	3.94	4.06	5.79	4.84	7.36	5.43
	5	7.63	7.71	9.44	8.72	12.13	9.59
	10	10.84	10.90	12.63	12.12	16.35	13.23
	100	22.55	22.58	24.30	24.54	32.12	26.53
0.7	1	4.79	4.83	5.74	4.96	6.12	6.36
	2	8.17	8.19	9.10	8.35	9.71	10.42
	5	13.92	13.94	14.85	14.14	15.89	17.35
	10	18.95	18.96	19.87	19.20	21.32	23.41
	100	37.33	37.34	38.25	37.72	41.35	45.58
0.9	1	15.93	15.94	16.22	15.94	16.26	20.63
	2	25.42	25.42	25.70	25.43	25.77	32.80
	5	41.64	41.64	41.92	41.64	42.03	53.59
	10	55.82	55.82	56.10	55.82	56.26	71.78
	100	107.69	107.69	107.97	107.70	108.30	138.29

It seems reasonable to expect that the performance of the asymptotically optimal allocations would be even better if we replaced the objective function in problem P_p with the one in (25), which incorporates C_i. In any case, this would not affect the asymptotic optimality and seems likely to improve performance at lower utilizations; notice that $C_i = \rho_i$ in (31). Using the exact value of C_i is possible if either interarrival times or demands are exponentially distributed. In the first case, we have

$$C_i = \frac{1 - \rho_i}{\lambda_i \phi'_{D^i}(\gamma_i) - 1},$$

(combine Theorems IX.2.3(a) and XII.5.3 of Asmussen), and in the second case we have $C_i = \rho_i$ (Asmussen, Theorem IX.1.3(c)).

In a separate numerical investigation we have observed that our asymptotically optimal capacity allocation rule gives excellent results when applied to the problem of minimizing costs subject to service-level constraints. However, this comparison is somewhat less interesting than that reported in Tables 1 and 2, because, as explained at the end of Section 4, heavy-traffic limits cannot in general be used to meet the constraints. The same is true of proportional allocations.

Interestingly, in the case of (31), problem P_δ yields an *optimal* capacity allocation when $\delta_i \equiv \delta$ for all i, for some δ. For in this case we may take $B_i \equiv 1$ in (9); and the resulting surrogate objective function

$$\sum_i \frac{h_i}{\gamma_i(r_i)} \log(1/\delta),$$

is minimized at the same point as the true cost

$$\sum_i \frac{h_i}{\gamma_i(r_i)} [\log(1/\delta) + \rho_i(\delta - 1)],$$

though the two functions are clearly not the same.

Table II
Asymmetric Costs

ρ	(p_1, p_2, p_3)	Optimal	Asy-Opt	Asy-Asy	Hvy-Opt	Hvy-Hvy	Proport
0.5	1 2 4	3.46	3.56	5.46	4.82	7.56	5.64
	1 10 100	8.90	8.94	11.04	12.17	17.50	14.25
	4 2 1	4.70	4.83	6.38	5.26	7.67	5.64
	100 10 1	13.61	13.67	15.04	13.91	17.87	14.25
	1 10 1	4.23	4.35	6.13	5.48	8.46	6.41
	10 1 10	7.98	8.06	9.71	9.11	12.48	9.82
0.7	1 2 4	7.15	7.17	8.20	7.50	9.21	10.77
	1 10 100	14.29	14.30	15.48	15.40	19.36	25.12
	4 2 1	9.54	9.56	10.36	9.62	10.75	10.77
	100 10 1	24.12	24.13	24.81	24.15	25.55	25.12
	1 10 1	8.28	8.31	9.25	8.49	10.25	12.05
	10 1 10	14.44	14.46	15.33	14.74	16.53	17.73
0.9	1 2 4	21.59	21.59	21.92	21.60	22.06	33.85
	1 10 100	37.28	37.29	37.67	37.33	38.39	76.90
	4 2 1	30.00	30.00	30.24	30.00	30.28	33.85
	100 10 1	73.48	73.48	73.68	73.48	73.74	76.90
	1 10 1	24.58	24.58	24.87	24.58	24.98	37.68
	10 1 10	43.18	43.18	43.45	43.19	43.60	54.73

APPENDIX
CONVEXITY OF γ^{-1}

In this appendix, we prove Proposition 2. We treat a generic γ_i and drop the subscript i.

Let ψ_X and ψ_D be logarithmic moment generating functions for X and D, as in (17). Equation (7), defining γ, can be rewritten as

$$\psi_D(\gamma) + \psi_X(-r\gamma) = 0\,. \tag{33}$$

Regardless of the distribution of X, $\psi_X(\theta)$ is zero at the origin, decreases strictly to $-\infty$ as $\theta \to -\infty$ and increases strictly as θ increases to the radius of convergence of ψ_X. Thus, there exists an inverse $\psi_X^{\leftarrow}$ whose domain includes the point $-\psi_D(\gamma) < 0$, and we may rearrange (33) to get

$$r = \frac{g(\gamma)}{\gamma} \triangleq \frac{-\psi_X^{\leftarrow}(-\psi_D(\gamma))}{\gamma}\,.$$

Since γ is a function of r, the function $\theta \mapsto g(\theta)/\theta$ is the inverse of $r \mapsto \gamma(r)$. Moreover, since $\gamma(r)$ is increasing (this follows from its definition), $g(\theta)/\theta$ is increasing as well. If we can show that $g(\theta)/\theta$ is convex, it will follow that $\gamma(r)$ is concave and hence that $1/\gamma(r)$ is convex. Thus, the rest of the proof is devoted to showing that $g(\theta)/\theta$ is convex. We proceed with the following result:

Lemma 5. *Suppose that, on its domain, a function g is equal to its Taylor series about the origin and that $g^{(n)}(0) \geq 0$, for all $n = 0, 1, 2, \ldots$. Then $g(x)/x$ is convex.*

The proof amounts to differentiating $g(x)/x$ twice and showing that the nonnegativity of the derivatives of g makes the second derivative of the ratio nonnegative. The details are straightforward so we omit them.

In light of Lemma 5, it suffices to show that all derivatives of $-\psi_X^{\leftarrow}(-\psi_D(\theta))$ at $\theta = 0$ are nonnegative, as the analyticity condition is automatically satisfied by logarithmic moment generating functions. This, in turn, is equivalent to showing that the derivatives of $\psi_X^{\leftarrow}(-\psi_D(\theta))$ are all less than or equal to zero. For this step we use the following lemma:

Lemma 6. *Suppose f and g are infinitely many times differentiable at the origin with $f(0) = g(0) = 0$, $g^{(n)}(0) \leq 0$ for all $n = 1, 2, \ldots$, and $f^{(n)}(0) \leq 0$ for odd n and $f^{(n)}(0) \geq 0$ for even n. Then $(f \circ g)^{(n)}(0) \leq 0$ for all n.*

This result is easily established by induction; we omit the details. To apply it to the problem at hand, notice that $\psi_D^{(n)}(0) \geq 0$ for all n; these derivatives are the *cumulant* moments of D and cannot be negative. Thus, $-\psi_D^{(n)}(0) \leq 0$ for all n. Under the hypothesis that the derivatives of $\psi_X^{\leftarrow}$ alternate signs, we conclude from Lemma 6 that the derivatives of $-\psi_X^{\leftarrow}(-\psi_D(\theta))$ are all nonnegative at $\theta = 0$, and hence that $g(\theta)/\theta$ is convex.

ACKNOWLEDGMENT

The author's research is supported by the National Science Foundation through grant MSS-9216490.

REFERENCES

AKELLA, R. AND P. R. KUMAR. 1986. Optimal Control of Production Rate in a Failure Prone Manufacturing System. *IEEE Trans. Automatic Cntrl.* **AC-31,** 116–126.

ANANTHARAM, V. 1989. The Optimal Buffer Allocation Problem. *IEEE Trans. Infor. Theory* **35,** 721–725.

ASMUSSEN, S. 1987. *Applied Probability and Queues*. Wiley, New York.

BREIMAN, L. 1961. Optimal Gambling Systems for Favorable Games. *Fourth Berkeley Symposium on Probability and Statistics, I*, 65–78.

CARR, S. A., A. R. GÜLLÜ, P. L. JACKSON, J. MUCKSTADT, AND R. ROUNDY. 1993. On the No B/C Stock Policy: A Partial Make-to-Order Strategy. Working Paper, School of ORIE, Cornell University, Ithaca, NY.

FEDERGRUEN, A. AND P. ZIPKIN. 1986. An Inventory Model with Limited Production Capacity and Uncertain Demands, I: The Average Cost Criterion. *Math. Oper. Res.* **11,** 193–207.

FELLER, W. 1971. *An Introduction to Probability Theory and its Applications*, Vol. 2., Second Edition, Wiley, New York.

GLASSERMAN, P. 1993. Bounds and Asymptotics for Planning Critical Safety Stocks. *Opns. Res.* to appear.

IBARAKI, T. AND N. KATOH. 1988. *Resource Allocation Problems: Algorithmic Approaches*. MIT Press, Cambridge, Massachusetts.

KINGMAN, J. F. C. 1970. Inequalities in the Theory of Queues. *J.R. Statist. Soc. B* **32,** 102–110.

KLEINROCK, L. 1976. *Queueing Systems, Vol. II*. Wiley-Interscience, New York.

LUSS, H. AND S. K. GUPTA. 1975. Allocation of Resources Among Competing Activities. *Oper. Res.* **23,** 360–366.

PRABHU, U. 1980. *Stochastic Storage Systems: Queues, Insurance Risk and Dams*. Springer, New York.

REIMAN, M. I. 1990. Some Allocation Problems for Critically Loaded Loss Systems with Independent Links. In *Performance '90*, P. J. B. King, I. Mitrani, and R. J. Pooley (eds.). Elsevier Science Publishers, New York.

ROSS, S. M. 1974. Bounds on the Delay Distribution in GI/G/1 Queues. *J. Appl. Prob.* **11,** 417–421.

SIEGMUND, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York.

TAYUR, S. 1993. Computing the Optimal Policy for Capacitated Inventory Models, *Stochastic Models*. to appear.

WEIN, L. M. 1989. Capacity Allocation in Generalized Jackson Networks. *Oper. Res. Lett.* **8,** 143–146.

WEIN, L. M. 1992. Dynamic Scheduling of a Multiclass Make-to-Stock Queue. *Oper. Res.* **40,** 724–735.

ZHENG, Y. S. AND P. ZIPKIN. 1990. A Queuing Model to Analyze the Value of Centralized Iventory Information. *Opns. Res.* **38,** 296–307.

ZIPKIN, P. 1980. Simple Ranking Methods for Allocating One Resource. *Mgmt. Sci.* **26,** 34–43.