M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 10, No. 1, Winter 2008, pp. 43–60
ISSN 1523-4614 | EISSN 1526-5498 | 08 | 1001 | 0043

informs

DOI 10.1287/msom.1070.0153

Demand Allocation in Systems with Multiple Inventory Locations and Multiple Demand Sources

Saif Benjaafar
Graduate Program in Industrial Engineering, Department of Mechanical Engineering, University of Minnesota, Minneapolis, Minnesota 55455, saif@umn.edu

Yanzhi Li
Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong, yanzhili@cityu.edu.hk

Dongsheng Xu
Department of Management Science, School of Business, Sun Yat-Sen University, Guangzhou, Guangdong 510275, China, xudongsheng@gmail.com

Samir Elhedhli
Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada, elhedhli@engmail.uwaterloo.ca

We consider the problem of allocating demand that originates from multiple sources among multiple inventory locations. Demand from each source arrives dynamically according to an independent Poisson process. The cost of fulfilling each order depends on both the source of the order and its fulfillment location. Inventory at all locations is replenished from a shared production facility with a finite production capacity and stochastic production times. Consequently, supply lead times are load dependent and affected by congestion at the production facility. Our objective is to determine an optimal demand allocation and optimal inventory levels at each location so that the sum of transportation, inventory, and backorder costs is minimized. We formulate the problem as a nonlinear optimization problem and characterize the structure of the optimal allocation policy. We show that the optimal demand allocations are always discrete, with demand from each source always fulfilled entirely from a single inventory location. We use this discreteness property to reformulate the problems as a mixed-integer linear program and provide an exact solution procedure. We show that this discreteness property extends to systems with other forms of supply processes. However, we also show that supply systems exist for which the property does not hold. Using numerical results, we examine the impact of different parameters and provide some managerial insights.

Key words: production-inventory systems; optimal demand allocation; make-to-stock queues; facility location
History: Received: October 17, 2005; accepted: December 21, 2006. Published online in *Articles in Advance* December 11, 2007.

1. Introduction

Global manufacturing firms are often faced with the need to consolidate manufacturing operations in a single low-cost location. However, to serve their customers effectively, they must also maintain multiple distribution centers from which to fulfill demand from different markets. The size and location of these distribution centers depend on which markets are served from which distribution center and the costs associated with operating a center at a particular location. An important trade-off for these firms is one between transportation and inventory costs. Operating few distribution centers allows a firm to pool inventory in few locations and therefore reduce risk from fluctuation in demand. Operating multiple distribution centers, on the other hand, reduces transportation costs by letting each market be served by the closest possible location. This trade-off is particularly important when transshipment between distribution centers is not feasible or not allowed and demand in each market is highly variable.

In this paper, we consider the problem of how a firm should allocate demand for a single product that originates from multiple sources (markets) among multiple inventory locations (distribution centers). The demand from each source occurs continuously

over time with stochastic interarrival times between individual orders. The cost of fulfilling each order depends on both the source of the order and its fulfillment location. We refer to this as a transportation cost, although it may correspond to other origin- or destination-sensitive costs. Each location can stock inventory in anticipation of future demand. However, it incurs a holding cost per unit of inventory per unit time, which may vary by location. If an order cannot be fulfilled immediately from inventory from its assigned location, it is backordered, but it incurs a backorder cost per unit of time the order is delayed. All locations are supplied from a shared production facility with a finite production rate and stochastic production times. Consequently, supply lead times from the production facility to the inventory locations are stochastic and affected by the congestion at the production facility. Inventory at each location is managed using a base-stock policy with a stationary base-stock level.

The objective is to determine (1) the fraction of demand from each source to allocate to each location and (2) the inventory level to keep at each location so that the sum of transportation, inventory holding, and demand backordering costs is minimized. Note that the demand allocation and inventory control problems must be solved jointly because the choice of optimal base-stock levels is affected by the demand allocation, and vice versa.

The optimal solution involves balancing two trade-offs. The first favors assigning demand from each source to the location with the lowest transportation cost. The second favors consolidating the fulfillment of demand in as few locations as possible to benefit from *inventory pooling* (by centralizing inventory in few locations, the probability of backordering can be reduced without increasing the investment in inventory). The relative strengths of these two trade-offs depend largely on the values of unit transportation, and holding and backordering costs. For example, if holding costs are negligible, large inventory amounts can be kept in each location, and it would be optimal to assign the demand from each source to the location with the lowest transportation cost for that source. On the other hand, if transportation costs are negligible, or are the same for all locations, then the optimal allocation would depend only on backorder and holding costs (if backorder costs are the same across locations, it would be optimal to assign all demand to the location with the lowest holding cost). Between these two extremes, it is not clear how demand should be allocated or how much inventory should be held in each location. For such cases, it may neither be desirable to satisfy each demand source from the closest inventory location nor to pool all inventory in a single location. It is also not clear if the demand from each source should be allocated in its entirety to a single location or be split among multiple locations.

In addition to the above direct effects of transportation and inventory costs, an indirect effect due to congestion at the production facility plays an equally important role. In systems where the utilization of the production facility is high, congestion is also high, and supply lead times are long. Consequently, the need for inventory at the various locations grows, increasing the desirability for inventory pooling. In contrast, when the utilization of the production facility is low, supply lead times are short, and there is less need for inventory, diminishing the benefit of inventory pooling and increasing the desirability of using the closest location.

The joint demand allocation and inventory control problem arises in a variety of settings. As mentioned earlier, the problem is faced by most manufacturing firms that manage multiple distribution warehouses with demand that originates from multiple geographical locations. The problem also arises in the context of a firm that produces multiple variants of the same component used for different products, with some variants serving as potential substitutes for others (Thonemann and Brandeau 2000). Despite its prevalence, in these and other settings, the problem has not been fully addressed in the literature.

There is, of course, a large body of literature dealing with the related problem of joint facility location and demand allocation (see, for example, Cornuéjols et al. 1990, Labbé and Louveaux 1997, Sherali et al. 2002, Daskin et al. 2005). However, most of that literature focuses on transportation-related costs in systems with deterministic demand and capacity. Shen et al. (2003) do consider a location model with inventory considerations. However, in their case, supply lead times are constant, and splitting the demand that originates from the same source among multiple locations is not allowed.

There is a rich literature dealing with inventory pooling (see the seminal paper by Eppen 1979 and recent papers by Gerchak and He 2003 and Benjaafar et al. 2005). This literature is primarily concerned with quantifying the benefits of pooling, assuming uniform or negligible transportation costs. There is related literature dealing with component commonality and substitution. See, for example, Gerchak and Henig (1989), Bassok et al. (1999), van Mieghem and Rudi (2002), and Netessine et al. (2002). Many of these papers consider a single-period problem in which decisions about inventory levels of each component are made prior to observing actual demand. Once the random demand is realized, an allocation is carried out either via a static allocation rule or by optimally solving an assignment problem. Benjaafar et al. (2004) treat a problem similar to ours with multiple products and multiple production facilities. However, in their case, demand for each product originates from a single source. Hence, their model does not include the trade-off from transportation costs that arises in ours.

Finally, there is related literature in queueing theory that deals with allocating demand among multiple servers, where the objective is to minimize a measure of customer delay or customer delay cost (a pure queueing system can be viewed as a make-to-order system where no inventory is held in anticipation of future demand). Examples from this literature include Bell and Stidham (1983), Tang and van Vliet (1994), Liu and Righter (1998), Benjaafar and Gupta (1999), and references therein. Several important cases are also discussed in Buzacott and Shanthikumar (1993). A closely related problem is the *load-sharing problem* that arises in the design of distributed computer systems. The literature on this topic is extensive, and examples include Wang and Morris (1985), Ni and Hwang (1985), and Bonomi and Kumar (1990). In §6.4, we comment more on the structure of optimal allocations that arise in these queueing settings.

To our knowledge, our paper is the first to consider the problem of joint demand allocation and inventory control in a system with continuous time, multiple demand sources, multiple inventory locations, and a capacitated production system. First, we consider systems with Poisson demand and exponentially distributed production times. We provide a model and exact solution procedure for determining the optimal demand allocations and optimal base-stock levels. We characterize the structure of the optimal allocation and show, perhaps surprisingly, that the optimal demand allocations are always discrete, so that it is always optimal to satisfy the entire demand from each source from a single inventory location. We extend our analysis to systems with other forms of supply processes and show that this discreteness property continues to hold in several cases. However, we also show that the discreteness property does not always hold. Indeed, systems with supply processes exist for which splitting demand can be desirable. Using numerical examples, we highlight the impact of various cost parameters on the optimal allocation and draw some insights.

The rest of this paper is organized as follows. In §2, we formulate the problem. In §3, we characterize the structure of the optimal allocation and use this structure to develop an exact solution procedure. In §4, we provide some insights from numerical results. In §§5, 6, and 7, we extend the analysis to systems with fixed location costs, alternative supply processes, and general demand distributions, respectively. In §8, we offer a summary and concluding comments.

2. Model Formulation

We consider a system consisting of a single product, m inventory locations, n sources of demand, and a single production facility. Product demand from each source i $(i = 1, \ldots, n)$ occurs continuously over time according to a Poisson process with rate λ_i. There is a cost c_{ij} of fulfilling an order of source i from location j $(j = 1, \ldots, m)$. This cost captures both the unit transportation cost from source i to location j and from the production facility to location j. Each location may hold inventory in anticipation of demand. However, there is a holding cost h_j per unit of inventory held in location j per unit time. If demand cannot be immediately satisfied from inventory, it is backordered, but there is a backorder cost b_j per unit backordered per unit time. Both unit holding and backorder costs may vary from location to location, so that, in general, we may have $h_i \neq h_j$ and $b_i \neq b_j$ for $i \neq j$. Inventory at each location is replenished from a single production facility shared among the different inventory locations. Orders from the different locations are processed at the production facility—which does not hold any inventory of its own—on a

first-come, first-served (FCFS) basis. Inventory at each location is managed using a base-stock policy with base-stock level s_j at location j. This means that the arrival of an order at a location triggers the placement of a replenishment order with the production facility. Production times at the facility are exponentially distributed with mean $1/\mu$, where, for stability, we assume $\sum_{i=1}^{n} \lambda_i < \mu$. Hence, the production system behaves like an $M/M/1$ queue (alternative models for the supply process are discussed in §6).

We consider two types of decisions: (1) the fraction α_{ij} of demand from source i assigned to location j, where $0 \leq \alpha_{ij} \leq 1$, and (2) the base-stock level s_j at each location j. The fraction α_{ij} can also be viewed as the probability that an order from source i is satisfied from location j. In practice, a truly probabilistic demand allocation is unlikely. However, it is useful in approximating the behavior of a central dispatcher that attempts to adhere to specified allocation ratios, or in modeling settings where demand from each source arises from a sufficiently large number of customers. For example, a distribution center may be responsible for fulfilling demand from a large number of retailers. In that case, the variable α_{ij} would correspond to the fraction of customers (e.g., retailers) from source i that is always satisfied from location j (as it turns out, a fractional allocation is never optimal, and the optimal allocations are always discrete; see §3).

Some assumptions are worth highlighting. First, we assume (unless stated otherwise) that orders are processed at the production facility on an FCFS basis. This is motivated by the widespread use of the FCFS policy in practice, its ease of implementation, perceived fairness, and analytical tractability. Characterizing an optimal policy for a system with multiple inventory-stocking locations and varying cost parameters is a difficult problem that remains unresolved; see de Véricourt et al. (2000) for results and references. The problem could be formulated as a Markov decision process (MDP) and solved numerically. However, such an approach is practically feasible only for small systems (e.g., with two locations) and for relatively low utilization levels at the production facility. Nevertheless, evidence shows that the difference in cost between the FCFS and an optimal policy diminishes as utilization increases, with this difference becoming negligible when utilization is high (see Zheng and Zipkin 1990, Zipkin 1995, van Houtum et al. 1997). Assigning static priorities among the different locations could provide an alternative to the FCFS policy. However, given the asymmetry in transportation, and backorder and holding costs, it is not clear how static priorities could be constructed. Furthermore, static priorities are analytically difficult to evaluate and can sometimes be less efficient than the FCFS policy; see de Véricourt et al. (2000) and Veatch and Wein (1996).

Second, we assume that transshipment of inventory from one location to another is not allowed. Therefore, our model applies only to settings where transshipments between inventory locations are prohibitively expensive or not feasible. This is not uncommon in practice. Consider, for example, a production facility that is centrally located (say in Hong Kong) but that supplies distribution centers in locations relatively distant from each other (e.g., Taipei, Tokyo, Beijing, and Bangkok). In this case, it may be cheaper to place orders directly with the production facility than to request a transshipment from another inventory location. Many firms also do not have the logistics to handle transshipments, which require a sophisticated information system and responsive transportation infrastructure. In terms of analysis, including transshipment adds considerable complexity. In fact, to our knowledge, the structure of the optimal policy in a setting like ours with transshipments is not known. We should note that not including transshipment in making initial demand allocations is consistent with standard models from location theory, including those that have attempted to account for the impact of inventory (see, for example, Shen et al. 2003).

Finally, we assume that a base-stock policy is used to manage inventory at each location. A base-stock policy is appropriate when ordering costs are not significant or when frequent deliveries are made from the production facility to the inventory locations, an increasingly common practice.

Our objective is to identify an allocation matrix $\boldsymbol{\alpha}^* = [\alpha_{ij}^*]$ and a base-stock-level vector $\mathbf{s}^* = (s_1^*, \ldots, s_m^*)$ that minimize the long-run expected total cost over all locations. Given an allocation matrix $\boldsymbol{\alpha}$ and a base-stock-level vector $\mathbf{s}$, expected total cost can be expressed as

$$z(\boldsymbol{\alpha}, \mathbf{s}) = \sum_{j=1}^{m} \left[h_j E(I_j) + b_j E(B_j) + \sum_{i=1}^{n} \alpha_{ij} c_{ij} \lambda_i \right], \qquad (1)$$

where I_j and B_j are random variables that denote inventory and backorder levels, respectively, at location j (note that both I_j and B_j depend on the choice of $\boldsymbol{\alpha}$ and $\mathbf{s}$). We refer to the above problem as the demand-allocation problem with distributed inventory (DAP-D).

Given an allocation matrix $\boldsymbol{\alpha}$ and a base-stock vector $\mathbf{s}$, expected inventory and backorder level can be obtained as follows (see, for example, Buzacott and Shanthikumar 1993, pp. 133–134):

$$E[I_j] = s_j - \left(1 - r_j^{s_j}\right) r_j/(1-r_j), \quad \text{and} \tag{2}$$

$$E[B_j] = r_j^{s_j+1}/(1-r_j), \tag{3}$$

where

$$r_j = \sum_{i=1}^{n} \alpha_{ij}\lambda_i \Big/ \left(\mu - \sum_{k\neq j}\sum_{i=1}^{n} \alpha_{ik}\lambda_k\right) = \hat{\lambda}_j/(\mu - \lambda + \hat{\lambda}_j), \tag{4}$$

with $\hat{\lambda}_j = \sum_{i=1}^{n} \alpha_{ij}\lambda_i$ corresponding to the overall demand rate assigned to location j and $\lambda = \sum_{i=1}^{n} \lambda_i$ corresponding to the aggregated demand flow from all sources. The joint demand allocation and inventory control problem can now be formulated as follows.

DAP-D:

$$\text{minimize} \quad z(\boldsymbol{\alpha}, \mathbf{s}) = \sum_{j=1}^{m}\left\{h_j\left[s_j - \left(\frac{r_j}{1-r_j}\right)\left(1 - r_j^{s_j}\right)\right] + b_j\left(\frac{r_j^{s_j+1}}{1-r_j}\right) + \sum_{i=1}^{n} c_{ij}\alpha_{ij}\lambda_i\right\} \tag{5}$$

$$\text{subject to} \quad \sum_{j=1}^{m} \alpha_{ij} = 1, \quad i = 1, \ldots, n; \tag{6}$$

$$\alpha_{ij} \geq 0, \quad i = 1, \ldots, n;\ j = 1, 2, \ldots, m; \tag{7}$$

$$s_j\text{: integer}, \quad j = 1, \ldots, m. \tag{8}$$

Given an allocation matrix $\boldsymbol{\alpha}$, $z(\boldsymbol{\alpha}, \mathbf{s})$ can be shown to be jointly convex in the s_j's. Noting that z is also separable in the s_j's, the optimal base-stock level s_j^* at each inventory location j can be obtained as the smallest integer that satisfies the constraint

$$z(\boldsymbol{\alpha}, \mathbf{s} + \mathbf{e}_j) - z(\boldsymbol{\alpha}, \mathbf{s}) \geq 0, \tag{9}$$

where $\mathbf{e}_j$ is the jth unit vector of dimension m. This leads to

$$s_j^* = \left\lfloor \frac{\ln[\omega_j]}{\ln[r_j]} \right\rfloor, \tag{10}$$

where the notation $\lfloor x \rfloor$ refers to the largest integer that is smaller than or equal to x and $\omega_j = h_j/(h_j + b_j)$. To simplify the analysis, we relax the integrality on s_j^* and let $s_j^* = \ln[\omega_j]/\ln[r_j]$. The amount of error introduced by this relaxation is relatively small, especially when s_j^* is large. The approximation is asymptotically exact when $r_j \to 1$ or, equivalently, when $\rho \to 1$ where $\rho = \lambda/\mu$ refers to the utilization of the production facility (see Appendix 1 for further supporting arguments). Note that relaxing the integrality of the base-stock level is in line with standard treatments in the inventory literature (Zipkin 2000) and in the analysis of *make-to-stock* queues (Buzacott and Shantikumar 1993, Wein 1992, Zipkin 1995).

Substituting s_j^* in the objective function, we can rewrite the optimal cost for a given allocation matrix $\boldsymbol{\alpha}$ as follows

$$z(\boldsymbol{\alpha}) = \sum_{j=1}^{m}\left\{h_j s_j^* + \sum_{i=1}^{n} c_{ij}\alpha_{ij}\lambda_i\right\} = \sum_{j=1}^{m}\left\{h_j \frac{\ln[\omega_j]}{\ln[r_j]} + \sum_{i=1}^{n} c_{ij}\alpha_{ij}\lambda_i\right\}. \tag{11}$$

Hence, the DAP-D can be reduced to a problem of finding the optimal allocation matrix $\boldsymbol{\alpha}^*$. Unfortunately, the total cost function in (11) is not jointly convex in the decision variables α_{ij}. This makes the DAP-D difficult to solve directly. In the next section, however, we show that the optimal allocation has a particular structure of which we can take advantage to construct an effective solution procedure. Specifically, we prove that the optimal allocations are always discrete, with the optimal values for the variables α_{ij} always being zero or one. We will show that this discreteness property can be used to transform the problem into a mixed-integer linear program that can be solved effectively.

3. The Structure of the Optimal Allocation

Let $f_j(\boldsymbol{\alpha})$ refer to the inventory cost contribution (sum of holding and backordering costs) due to location j $(j = 1, 2, \ldots, m)$ given an allocation matrix $\boldsymbol{\alpha}$ and an optimal base-stock vector $\mathbf{s}^*$ for that allocation. From (11), we have $f_j(\boldsymbol{\alpha}) = h_j \ln[\omega_j]/\ln[r_j]$, from which we can see that $f_j(\boldsymbol{\alpha})$ depends on the allocation variables only via the sum $\hat{\lambda}_j = \sum_{i=1}^{n} \alpha_{ij}\lambda_i$, the overall

demand rate assigned to location j. In other words, if the demand rate $\hat{\lambda}_j$ is the same under two different allocations $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, then we have $f_j(\boldsymbol{\alpha}) = f_j(\boldsymbol{\alpha}')$. In the remainder of this section, we highlight this by referring to the function $f_j(\boldsymbol{\alpha})$ simply as $f_j(\hat{\lambda}_j)$.

THEOREM 1. *The inventory cost function, $f_j(\hat{\lambda}_j)$, at each location j is strictly concave in $\hat{\lambda}_j$. Therefore, the optimal demand allocations are always discrete with, $\alpha_{ij}^* = 0$ or 1 for all values of i and j.*

PROOF. It is straightforward to show that the function f_j is strictly concave by showing that the second derivative is negative. The proof that this leads to discrete allocations is as follows. Suppose a demand source k exists whose demand is split among a subset S of locations. Then, it is always possible to find a lower cost allocation that assigns all of the demand from that source to only one of the locations in the set S. More specifically, let u and w be two locations in the set S such that $0 < \alpha_{ku} < 1$ and $0 < \alpha_{kw} < 1$. Let $\lambda_u^- = \sum_{i \neq k} \alpha_{iu}\lambda_i$ and $\lambda_w^- = \sum_{i \neq k} \alpha_{iw}\lambda_i$ where λ_u^- and λ_w^- represent the demand assigned to locations u and w, respectively, from sources other than source k, and $\lambda_{uw} = \alpha_{ku}\lambda_k + \alpha_{kw}\lambda_k$ denote the demand from source k that is split between locations u and w, and let β be the fraction of λ_{uw} assigned to location u. The contribution to total cost from locations u and w, which we denote by $z_{uw}(\beta)$, can then be written as

$$z_{uw}(\beta) = f_u(\lambda_u^- + \beta\lambda_{uw}) + f_w(\lambda_w^- + (1-\beta)\lambda_{uw}) + \sum_{i \neq k} (\alpha_{iu}c_{iu} + \alpha_{iw}c_{iw})\lambda_i + (\beta c_{ku} + (1-\beta)c_{kw})\lambda_{uw}.$$

We can see that $z_{uw}(\beta)$ is strictly concave in β. Therefore, $z_{uw}(\beta)$ is minimum when $\beta = 0$ or $\beta = 1$ (the minimum of a strictly concave function occurs at an extreme point). This means that assigning the demand from source k split between the pair of locations u and w to either one of the two locations always reduces the cost contribution from u and w and, therefore, the total cost (note that the reallocation of demand between u and w does not affect the cost contribution from other locations). Starting from this new reallocation, we can apply a similar logic to any remaining pair of locations in S with fractional allocations of demand from source k to show that consolidating demand in one of the two locations is always optimal. Successive application of this procedure to all pairs of locations and to all demand sources yields an allocation in which the entire demand from each source is assigned to a single location. Hence, given a nondiscrete allocation of demand among locations, it is always possible to a find a discrete allocation with a lower cost. Consequently, the optimal allocation must be discrete. □

Theorem 1 is a general result that provides a sufficient condition (strict concavity of the inventory cost function) for the optimal allocation to be guaranteed to be discrete. The condition applies to any inventory system, with a properly redefined function f_j, regardless of its supply process; see §6 for examples. Note that although strict concavity of f_j guarantees the optimal allocation to be discrete, simple concavity is sufficient for the existence of a discrete optimal allocation.

In §6, we show that the discreteness property of the optimal allocation indeed holds in several additional settings. However, we also show that it is not always true and that systems with supply processes exist for which demand splitting can be desirable. The discreteness property can be used to transform the problem into an integer optimization problem. In Appendix 2, we describe such a reformulation and show how the reformulated problem can be solved effectively via a linearization procedure and a *constraints generation* algorithm. We also provide numerical results (see Table 3) that illustrate the computational effectiveness of the solution procedure.

4. Some Insights from Numerical Examples

Theorem 1 states that it is never optimal to split the demand from any source among multiple locations. This is perhaps surprising because the objective function contains costs with potentially counteracting effects. For example, facilities with the lowest transportation costs can be different from those with the lowest holding or backorder costs. Although the allocations are always discrete, we have found that they can be far from obvious. As we show in the following observation, it can be optimal not to assign demand to a location even when that location offers the lowest transportation, and holding and backorder costs.

OBSERVATION 1. It can be optimal to assign the demand from a source to a location with the highest

transportation, and inventory holding and backorder costs.

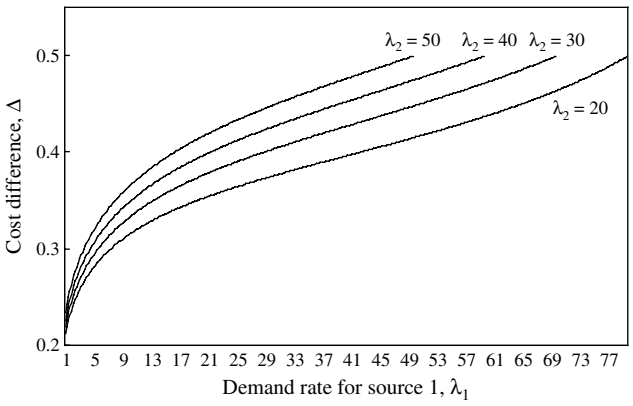
Observation 1 can be proven using the following example. Consider a system with three demand sources and two inventory locations with the following operating parameters: $c_{11} = 0.1$, $c_{12} = +\infty$, $c_{21} = 0.01$, $c_{22} = 0.015$, $c_{31} = +\infty$, $c_{32} = 0.1$, $h_1 = 1.0$, $h_2 = 1.05$, $b_1 = b_2 = 10$, $\lambda_1 = 1$, $\lambda_2 = 10$, and $\mu = 30$. First, it is easy to verify that demand from source 1 would always be allocated to location 1, and demand from source 3 would always be allocated to location 2. For source 2, one might expect that it would be optimal to allocate its demand to location 1 since location 1 has at the same time lower transportation, holding, and backorder costs. However, depending on the value of λ_3, it turns out that it can be optimal to assign the demand to either location 1 or 2. In particular, if λ_3 is in the range [2.3, 14.9], then it is optimal to allocate demand from source 2 to location 2; otherwise, it is optimal to allocate it to location 1.

The above example illustrates the subtle impact of inventory pooling and its sensitivity to the amount of demand at each facility. It is interesting to note that the effect of demand rate at each source on the optimal allocation is not monotonic. In the above example, when λ_3 is small, it is optimal to allocate demand from source 2 to location 1; when λ_3 is in the midrange it is optimal to allocate this demand to location 2; and when demand is sufficiently high, it is once again optimal to assign it to location 1. This behavior is in part due to the fact that the benefit of inventory pooling can exhibit diminishing returns as demand increases and appears to remain bounded.

Observation 2. The marginal benefit from inventory pooling can diminish with increases in the demand at one or more of the sources being pooled. Moreover, the absolute benefit from pooling tends to remain bounded with increases in the demand rates of these sources.

To illustrate the above observation, consider a system with two demand sources with rates λ_1 and λ_2 and two locations, 1 and 2. For simplicity, let all cost parameters be identical at the two locations. Let Δ refer to the difference between the optimal cost when the demand from each source is satisfied from a separate location (source 1 from location 1 and source 2 from location 2) and the optimal cost when both demand sources are satisfied from a single location. The value of Δ is shown in Figure 1 for different values of λ_1 and λ_2. We can see that the marginal increase in Δ tends to decrease (although not always) as λ_1 increases for fixed λ_2. More significantly, as shown below, Δ remains bounded as λ_1 increases and approaches its maximum feasible value of $\mu - \lambda_2$:

Figure 1 The Impact of Demand Rates on the Benefit of Inventory Pooling



$$\lim_{\lambda_1 \to (\mu - \lambda_2)} \Delta = \lim_{\lambda_1 \to (\mu - \lambda_2)} h \ln[h/(h+b)] \cdot \left\{ \frac{1}{\ln[(\lambda_1 + \lambda_2)/\mu]} - \frac{1}{\ln[\lambda_1/(\mu - \lambda_2)]} - \frac{1}{\ln[\lambda_2/(\mu - \lambda_1)]} \right\}, \quad (12)$$

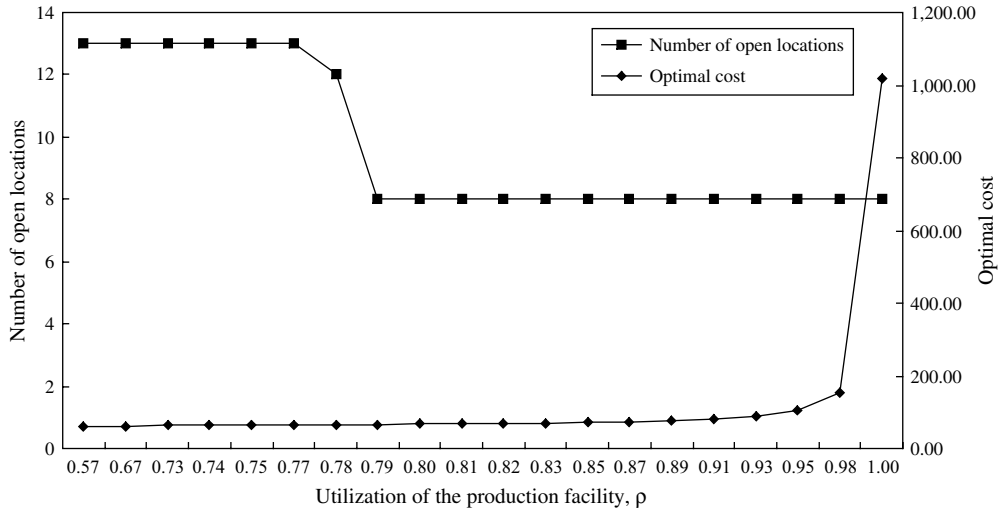
which, upon application of l'Hopital's rule, leads to

$$\lim_{\lambda_1 \to (\mu - \lambda_2)} \Delta = h \ln[h/(h+b)]/2. \quad (13)$$

In addition to being affected by the individual demand rates, the benefit of pooling is also sensitive to the expected production time, and, more generally, to the utilization of the production facility. We observe that the benefit of pooling tends to increase as utilization increases. This makes intuitive sense because higher utilization leads to more congestion at the production facility and longer, more variable supply lead times at the inventory locations. Consequently, as utilization increases, consolidating demand in few locations becomes more desirable. This means that the number of locations with positive demand allocation also tends to decrease.

Observation 3. The number of locations with positive demand allocation, or open locations, generally

Figure 2 The Impact of Production Facility Utilization on the Number of Open Locations



Note. Forty demand sources each with demand rate $\lambda = 10$; 40 potential inventory locations with identical holding and backorder costs $h = 1$ and $b = 10$; and transportation costs $c_{ij} = \sqrt{|i-j|}/50 + 0.1$ for all i and j.

decreases with increases in the utilization of the production facility.

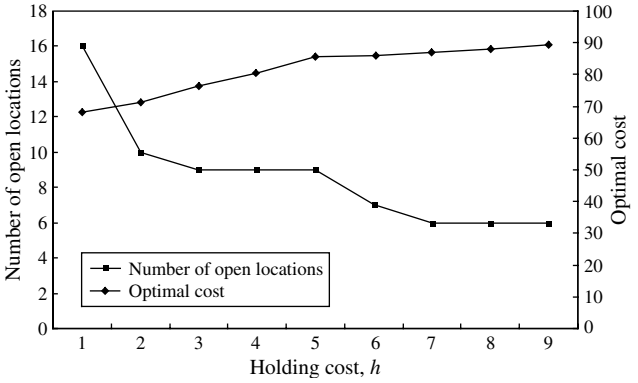
Observation 3 is illustrated in Figure 2 for an example system in which the utilization of the production facility is varied by changing the expected production time. As we can see, the utilization level can have a significant impact on how many locations are used. Interestingly, beyond a certain level, increases in utilization seem to have no effect on the number of open locations. This appears to be because the benefit of pooling remains bounded as utilization increases. Note that in our analysis, although we ignore the integrality of the base-stock levels, we expect the results to remain qualitatively the same when the integrality is enforced. When utilization is low, the dominant effect is that of transportation costs (because inventory levels are low). Therefore, demand would be allocated to the closest location, leading to a large number of locations being open. On the other hand, when utilization is high, the dominant effect is that of inventory costs. Therefore, it becomes desirable to consolidate demand in fewer locations.

Other factors that tend to affect the benefit of pooling include holding and backorder costs. This is illustrated for an example system in Figure 3, in which the number of open locations decreases with increases in holding cost.

The results shown in Figures 2 and 3 illustrate how the solution obtained from the DAP-D can be significantly different from a solution obtained using only transportation costs. They also highlight the complex interactions between transportation and inventory costs and production capacity and their effect on the optimal demand allocations. This leads to our final and most important observation.

Observation 4. The solution obtained from the DAP-D, and the corresponding optimal cost, can be

Figure 3 The Impact of Unit Holding Cost on the Number of Open Locations



Notes. Fifty demand sources each with demand rate $\lambda = 10$; $\mu = 800$. Fifty potential inventory locations, and transportation costs $c_{ij} = |i-j|/50 + 0.1$ for all i and j.

significantly different from those obtained using models that account for transportation costs but not for production capacity or inventory-related costs.

5. Systems with Fixed Location Costs

We have assumed that no fixed cost exists for having an inventory location. This is appropriate when the locations already exist and there is only an operational decision about demand allocation. However, in settings where we must decide about whether to invest in each location, there is generally a fixed cost K_j associated with assigning demand to a particular location j. To include such costs in the DAP-D formulation, we need to introduce a new decision variable y_j, which takes the value of one if location j is assigned positive demand and the value of zero otherwise. Then, the original demand allocation problem can be reformulated as follows:

$$\text{minimize} \quad z(\boldsymbol{\alpha}, \mathbf{s}) = \sum_{j=1}^{m} K_j y_j + \sum_{j=1}^{m} \left\{ h_j \left[s_j - \left(\frac{r_j}{1-r_j} \right) \left(1 - r_j^{s_j}\right) \right] + b_j \left(\frac{r_j^{s_j+1}}{1-r_j} \right) + \sum_{i=1}^{n} c_{ij} \alpha_{ij} \lambda_i \right\} \quad (14)$$

$$\text{subject to} \quad \sum_{j=1}^{m} \alpha_{ij} = 1, \quad i = 1, \ldots, n; \quad (15)$$

$$\alpha_{ij} - y_j \le 0, \quad i = 1, \ldots, n;\ j = 1, \ldots, m; \quad (16)$$

$$\alpha_{ij} \ge 0, \quad i = 1, \ldots, n;\ j = 1, \ldots, m; \quad (17)$$

$$s_j\text{: integer}, \quad j = 1, \ldots, m; \quad (18)$$

$$y_j \in \{0, 1\}, \quad j = 1, \ldots, m. \quad (19)$$

The above problem generalizes the classical uncapacitated location problem in which only transportation costs are considered (our problem reduces to a pure location problem when either $h = 0$, $b = 0$, or $\rho = 0$). An analysis similar to the one in §3 can be used to show that the optimal demand allocations remain discrete. Hence, the problem can be solved as an integer optimization problem. The procedure described in Appendix 2 applies to this case as well and leads to an efficient solution approach. Numerical results are provided in Table 4. Note that, qualitatively, the presence of fixed location costs tends to further favor pooling. Consequently, the optimal number of locations would tend to decrease with increases in the fixed location costs.

6. Extensions to Systems with Other Supply Processes

In this section, we extend the analysis to systems with alternative supply processes. We present models for each case and examine the extent to which, either exactly or approximately, the discrete allocation property continues to hold. We focus on the original problem without fixed location costs. However, all the results remain valid for problems with fixed location costs.

6.1. Systems with General Production Time Distributions

In some settings, it may be difficult to justify using the exponential distribution to describe production times. In those cases, a more appropriate model would allow production times to have a general distribution. This means that the production system would behave like an $M/G/1$ queue instead of an $M/M/1$ queue. Unfortunately, characterizing analytically the distribution of queue size in an $M/G/1$ queue—necessary for obtaining closed-form expressions for expected inventory and backorder levels—is difficult. A commonly used alternative is to approximate the distribution of queue size by a geometric distribution with a matching first moment (see, for example, Buzacott and Shantikumar 1993 and Tijms 1995 for supporting arguments and discussion). This leads to the following approximate expressions for expected inventory and backorder levels:

$$E[I_j] = s_j - \frac{\rho}{\sigma} \left(\frac{\hat{r}_j}{1-\hat{r}_j} \right) \left(1 - \hat{r}_j^{s_j}\right), \quad \text{and} \quad (20)$$

$$E[B_j] = \frac{\rho}{\sigma} \left(\frac{\hat{r}_j^{s_j+1}}{1-\hat{r}_j} \right), \quad (21)$$

where $\hat{r}_j = \hat{\lambda}_j \sigma / [\mu - \sigma(\lambda - \hat{\lambda}_j)]$, $\sigma = (E(Q) - \rho)/E(Q)$, and $E(Q) = [\lambda E[S^2]/2(1-\rho)] + \rho$, with $E(Q)$ representing the *exact* expected queue size (number of items

in queue + in service) in an $M/G/1$ queue and S a random variable that denotes production time.

Given an allocation matrix $\boldsymbol{\alpha}$, the optimal base-stock level at each inventory location j is given by $s_j^* = \ln[h_j\sigma/(\rho(h_j+b_j))]/\ln[\hat{r}_j]$, which, upon substitution into the objective function, leads to

$$z(\boldsymbol{\alpha}) = \sum_{j=1}^{m}\left\{h_j\frac{\ln[\sigma h_j/\rho(h_j+b_j)]}{\ln[\hat{r}_j]} + \sum_{i=1}^{n} c_{ij}\alpha_{ij}\lambda_i\right\}. \quad (22)$$

The inventory cost contribution (sum of holding and backordering costs) due to location j is given by

$$f_j(\hat{\lambda}_j) = \frac{\ln[\sigma h_j/\rho(h_j+b_j)]}{\ln[\hat{\lambda}_j\sigma/[\mu-\sigma(\lambda-\hat{\lambda}_j)]]}, \quad (23)$$

which can be verified to be strictly concave. Consequently, the optimal demand allocations are discrete with $\alpha_{ij}^* = 0$ or 1 for all values of i and j. The procedure described in Appendix 2 can be adapted to reformulate and solve the problem as an MILP.

As we can see from Equation (23), inventory costs are increasing, via the parameter σ, in the variability of production times. As variability increases, both the mean and variance of supply lead times increase. Consequently, the benefit of inventory pooling tends to increase as variability increases.

6.2. Systems with Exogenous and Sequential Supply Lead Times

Instead of explicitly modeling the supply process via a shared production facility with supply lead times that are load dependent, it is sometimes appropriate to model replenishment lead times as being exogenous. For example, this may be the case when the production facility and the inventory locations are owned by independent firms, and the load contributed by a particular location is insignificant relative to the total load of the facility. It may also be appropriate when orders are almost instantaneously fulfilled by the production facility (because of either ample capacity or inventory), but transportation lead times from the production facility to the inventory locations are significant. Additional discussion of exogenous and sequential supply lead times can be found in Zipkin (2000, pp. 273–279).

Let L_j be a random variable that denotes the exogenous and sequential supply lead times at location j ($j = 1, 2, \ldots, m$). To obtain expressions for expected inventory and backorder levels, we need to characterize the distribution of Q_j, which now has the interpretation of *inventory-on-order* at location j (i.e., the total number of orders that have been placed by location j, but that have not yet been delivered). Unfortunately, obtaining a closed-form expression for the distribution of Q_j is difficult in general. However, for the special case where L_j has the exponential distribution with mean $1/\mu_j$, we can show that Q_j has the geometric distribution with parameters $p_j = \hat{\lambda}_j/(\mu_j+\hat{\lambda}_j)$. For a given allocation matrix $\boldsymbol{\alpha}$ and base-stock vector $\mathbf{s}$, expected inventory and backorder levels are then given by

$$E[I_j] = s_j - (1-p_j^{s_j})p_j/(1-p_j), \quad \text{and} \quad (24)$$

$$E[B_j] = p_j^{s_j+1}/(1-p_j). \quad (25)$$

A development similar to the one described in §§2 and 3 can be used to formulate and solve the demand allocation problem. The cost contribution due to location j can be shown to be given by

$$f_j(\hat{\lambda}_j) = \frac{h_j\ln[\omega_j]}{\ln[\hat{\lambda}_j] - \ln[\mu_j+\hat{\lambda}_j]}, \quad (26)$$

which is strictly concave in $\hat{\lambda}_j$. Consequently, the optimal demand allocation remains discrete with $\alpha_{ij}^* = 0$ or 1.

For supply lead times with a general distribution, it is difficult to obtain an explicit characterization of Q_j. However, we can take advantage of the following relationship between the z-transform of Q_j, $\hat{g}_{Q_j}$, and the Laplace transform of the supply lead time L_j, f_{L_j} (see Chapter 7 of Zipkin 2000):

$$\hat{g}_{Q_j}(z) = f_{L_j}[\hat{\lambda}_j(1-z)], \quad (27)$$

from which we can derive the mean and variance of Q_j as $E(Q_j) = E(L_j)\hat{\lambda}_j = \hat{\lambda}_j/\mu_j$ and $\mathrm{Var}(Q_j) = \mathrm{Var}(L_j)\hat{\lambda}_j^2 + E(L_j)\hat{\lambda}_j = \hat{\lambda}_j^2/\mu_j^2 + \hat{\lambda}_j/\mu_j$, respectively, where $E(L_j)$ and $\mathrm{Var}(L_j)$ are the mean and variance of L_j and $\mu_j = 1/E[L_j]$, respectively.

If we approximate the distribution of Q_j by a normal distribution with matching mean $E(Q_j)$ and variance $\mathrm{Var}(Q_j)$ (see Chapters 6 and 7 of Zipkin 2000 for an extensive discussion of the appropriateness of the *normal approximation* for this and for other inventory

contexts), then expected total cost for a given allocation matrix $\boldsymbol{\alpha}$ and base-stock vector $\mathbf{s}$ is given by

$$z(\boldsymbol{\alpha}, \mathbf{s}) = \sum_{j=1}^{m} \Big\{ h_j E[(s_j - Q_j)^+] + b_j E[(Q_j - s_j)^+] + \sum_{i=1}^{n} c_{ij}\alpha_{ij}\lambda_i \Big\}. \quad (28)$$

Using the first-order condition of optimality, we can show that the optimal base-stock level at each location j is given by $s_j^* = E[Q_j] + z_j^*\sqrt{\mathrm{Var}(Q_j)}$, where z_j^* satisfies $\Phi(z_j^*) = b_j/(b_j + h_j)$ and Φ is the cumulative density function of the standard normal distribution. Substituting s_j^* into the objective function, we can obtain

$$z(\boldsymbol{\alpha}) = \sum_{j=1}^{m} \Big\{ (h_j + b_j)\sqrt{\mathrm{Var}(Q_j)}\phi(z_j^*) + \sum_{i=1}^{n} c_{ij}\alpha_{ij}\lambda_i \Big\}, \quad (29)$$

from which we can in turn obtain $f_j(\hat{\lambda}_j) = (h_j + b_j) \cdot \sqrt{\mathrm{Var}(Q_j)}\phi(z_j^*)$, where ϕ is the probability density function of the standard normal distribution. We can easily check that $\sqrt{\mathrm{Var}(Q_j)} = \sqrt{\mathrm{Var}(L_j)\hat{\lambda}_j^2 + E(L_j)\hat{\lambda}_j}$ is strictly concave in $\hat{\lambda}_j$ and, therefore, so is f_j. Thus, the optimal demand allocation remains discrete in this case as well.

6.3. Systems with Independent and Identically Distributed Supply Lead Times

In this section, we consider another commonly used model of the supply process, one in which supply lead times are exogenous, but independent and identically distributed (*i.i.d.*). The main difference between this model and the one in §6.2 is that here orders are not necessarily delivered in the sequence in which they have been placed (i.e., the FCFS assumption no longer holds in this case). Instead, the supply lead times are independent.

In systems in which demand occurs according to a Poisson process, it can be shown that the inventory on order at each location Q_j, $j = 1, \ldots, m$, also has the Poisson distribution. In particular, if the expected value of supply lead times is $1/\mu$, then Q_j also has the Poisson distribution with expected value $\hat{\lambda}_j/\mu$. Because it is difficult to obtain closed-form expressions for performance measures of interest, it is not uncommon to approximate the distribution of Q_j by a normal distribution with matching mean and variance. This approximation is particularly effective when $\hat{\lambda}_j$ is large. Under the normal approximation, an analysis similar to the one described in §6.2 leads to the following expression for $f_j(\hat{\lambda}_j)$:

$$f_j(\hat{\lambda}_j) = (h_j + b_j)\sqrt{\mathrm{Var}(Q_j)}\phi(z_j^*) = (h_j + b_j)\sqrt{\hat{\lambda}_j/\mu}\,\phi(z_j^*). \quad (30)$$

It can once again be shown that $f_j(\hat{\lambda}_j)$ is strictly concave in $\hat{\lambda}_j$. Hence, the optimal demand allocations are discrete.

6.4. Systems with Independent Capacitated Production Facilities

The results of §§6.1–6.3 could lead us to believe that the discreteness of the optimal allocation holds universally. In this section, we show that this is not true, and that in some settings a fractional allocation of the demand from the same source among multiple locations can be optimal. Consider a system identical to the one in §2, except that instead of a single production facility shared among the different locations, each location has its own independent production facility from which the inventory for that location is exclusively replenished. Furthermore, production times at each facility are exponentially distributed with mean $1/\mu_j$ for facility j. Hence, each facility behaves like an $M/M/1$ queue with arrival rate $\hat{\lambda}_j$, and expected inventory and backorder levels at each location are given by

$$E(I_j) = s_j - (1 - \rho_j^{s_j})\rho_j/(1 - \rho_j), \quad \text{and} \quad (31)$$

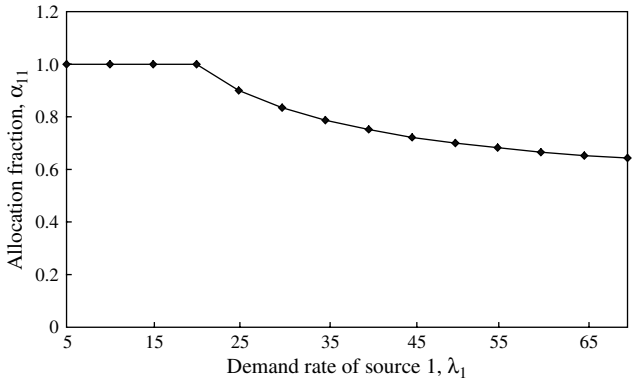
$$E(B_j) = \rho_j^{s_j+1}/(1 - \rho_j), \quad (32)$$

where $\rho_j = \hat{\lambda}_j/\mu_j$ corresponds to the utilization of the production facility at location j. The optimal base-stock level at location j is given by $s_j^* = \ln[\omega_j]/\ln[\rho_j]$, which, upon substitution into the objective function, leads to

$$z(\boldsymbol{\alpha}) = \sum_{j=1}^{m} \Big\{ h_j \frac{\ln[\omega_j]}{\ln[\rho_j]} + \sum_{i=1}^{n} c_{ij}\alpha_{ij}\lambda_i \Big\}. \quad (33)$$

The inventory cost contribution due to location j is given by $f_j(\hat{\lambda}_j) = h_j(\ln[\omega_j]/\ln[\hat{\lambda}_j/\mu_j])$, which, upon

Figure 4 Optimal Allocation from Source 1 to Location 1



Note. Two locations and two demand resources; $h_1 = h_2 = 1$, $b_1 = b_2 = 10$, $\mu_1 = \mu_2 = 50$, $c_{11} = c_{12} = c_{22} = 0.1$, $c_{21} = \infty$, $\lambda_2 = 20$.

examination, can be easily seen not to be concave in $\hat{\lambda}_j$, but convex instead. Consequently, the optimal demand allocations are not guaranteed to be discrete. In fact, it is not difficult to construct examples where a fractional allocation is optimal. This makes intuitive sense because expected queue size (and, consequently, expected supply lead time) at each facility increases in a convex fashion with increases in the workload of each facility. There is a need to balance the workload among the different production facilities, which in turn could make demand splitting desirable. This result is illustrated in Figure 4, where we show how the optimal allocation to location 1 in a system with two locations and two demand sources is affected by the demand rate from source 1. The special case in which inventory cannot be held (e.g., $h = \infty$) and no transportation costs exist has been studied by Bell and Stidham (1983), who provide closed-form expressions for the optimal allocations. In this case, it is generally optimal to split demand among multiple facilities (e.g., in the case of identical facilities, it is optimal to evenly split demand among all facilities).

7. Extensions to Systems with General Demand Processes

The approximation approach used in §6.1 to model systems with general production time distributions can in principle be used to model systems in which the demands also have general distributions but still form independent renewal processes. This would allow us to model the production facility as a $GI/G/1$ queue. Two difficulties, however, arise: (1) the superposition of renewal processes does not necessarily produce a renewal process, and therefore, the arrival process to the production facility may not be a renewal process; and (2) there are no known exact expressions for expected queue size in a $GI/G/1$ queue. The first difficulty may be handled by approximating superposed renewal processes by a renewal process whose coefficient of variation is obtained via a two-moment approximation; see, for example, Albin (1984) and Whitt (1982). The second difficulty can be addressed by using one of the many reasonably good approximations of expected queue size in a $GI/G/1$ queue; see, for example, Wolff (1989), Whitt (1983, 1993), and Buzacott and Shanthikumar (1993). Alternatively, we may focus on regimes in which explicit results are available. One such regime is heavy traffic. In particular, it is known that as the number of class j customers (orders due to the inventory location j in our case) in a multiclass $GI/G/1$ queue j weakly converges to a reflected Brownian motion with drift $\hat{\lambda}_j\rho^{-1}(1-\rho)$ and variance (Peterson 1991):

$$\hat{\lambda}_j\rho^{-2}\sum_{j=1}^{m}\hat{\lambda}_j E[S]^2\left(\widehat{C}_{a_j}^2 + C_S^2\right), \tag{34}$$

where $\widehat{C}_{a_j}$ is the coefficients of variation in interarrival times to inventory location j. Let C_a refer to the coefficients of variation in order interarrival times to the production facility and C_{a_i} to coefficient of variation of interarrival times to demand source i, respectively. Then, using the approximation (Whitt 1982):

$$C_a^2 = \sum_{j=1}^{m}\frac{\hat{\lambda}_j}{\lambda}\widehat{C}_{a_j}^2 \tag{35}$$

leads to

$$\sum_{j=1}^{m}\hat{\lambda}_j\left(\widehat{C}_{a_j}^2 + C_S^2\right) = \lambda(C_a^2 + C_S^2) = \sum_{i=1}^{n}\lambda_i\left(C_{a_i}^2 + C_S^2\right). \tag{36}$$

For a given demand allocation matrix, $\boldsymbol{\alpha}$, it can then be shown that the optimal base-stock level at location j is given by (see Wein 1992 for details):

$$s_j^* = \frac{\hat{\lambda}_j \ln[(h_j + b_j)/h_j]\sum_{i=1}^{n}\lambda_i\left(C_{a_i}^2 + C_S^2\right)}{2\mu(\mu - \lambda)}, \tag{37}$$

which, upon substitution in the objective function, leads to

$$z(\alpha, s^*) = \sum_{j=1}^{m} \left\{ \frac{\hat{\lambda}_j \ln[(h_j + b_j)/h_j] \sum_{i=1}^{n} \lambda_i \left(C_{a_i}^2 + C_S^2\right)}{2\mu(\mu - \lambda)} \right\} + \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \alpha_{ij} \lambda_i. \quad (38)$$

We can now see that the inventory cost contribution due to location j is given by

$$f_j(\hat{\lambda}_j) = \frac{\hat{\lambda}_j h_j \ln[(h_j + b_j)/h_j] \sum_{i=1}^{n} \lambda_i (C_{a_i}^2 + C_S^2)}{2\mu(\mu - \lambda)}, \quad (39)$$

which is linear in $\hat{\lambda}_j$. Hence, the discreteness of the optimal demand allocations (under the above approximations) continues to hold.

Note that inventory cost is an increasing function of variability, via the parameters C_{a_i} in the demand process at each source i, $i = 1, \ldots, n$. Hence, the benefit from inventory pooling tends to increase as demand variability increases. In turn, this can lead to fewer inventory locations being opened as variability increases.

8. Conclusions

In this paper, we have considered the problem of jointly allocating demand and determining optimal inventory levels in a system consisting of multiple inventory locations, multiple sources of demand, and a capacitated production process. For systems with Poisson demand and exponentially distributed production times, we formulated the problem as a nonlinear optimization problem and showed that the optimal demand allocations are always discrete, with demand from each source always fulfilled entirely from a single location. We found that the discreteness property would hold, regardless of the type of supply process, if the optimal inventory cost at each location is concave in the demand that is allocated to that location. We used this discreteness property to reformulate the problems as a mixed-integer linear program, and provided an exact solution procedure. We showed that the discreteness property extends to several other supply systems. However, we also showed that supply systems exist for which the concavity property, and the resulting discreteness, does not hold. Using numerical results, we examined the impact of various parameters.

There are several potential avenues for future research. It would be of interest to extend the analysis to systems with fixed ordering costs at the inventory locations or fixed setup costs/setup times at the production facility. In the case of fixed ordering costs, it would become desirable to place orders in batches. This means that more inventory would be held, making inventory pooling more attractive. Similarly, in the case of fixed setup costs/setup times at the production facility, it would become desirable to produce in batches, making production lead times longer and increasing the need for more inventory. In turn, this would make inventory pooling more useful. In short, the potential of economies of scale either in ordering or production is likely to strengthen the pooling effect, leading to fewer inventory locations and less likelihood of demand splitting. Another important future research avenue might be to consider systems in which backorder costs vary by both location and demand source, which would require that orders from different demand streams be treated differently. In particular, it could become optimal to *ration* inventory among the different demand classes by, for example, reserving inventory for future orders from classes with high backorder costs when inventory drops below certain thresholds. Finally, it would be of interest to relax some of our modeling assumptions by allowing orders to vary in size, to be correlated across locations, or not to be stationary over time. It would also be useful to develop solution procedures that directly treat inventory decision variables as discrete by taking advantage of possible problem structure such as convexity (see Murota 2003 for advances in discrete convex optimization) or by exploring potential linearization of the objective function and using standard approaches for linear mixed-integer programming.

Acknowledgments

The research of the first author was in part carried out while he was on sabbatical leave at Hong Kong University of Science and Technology. The first author is grateful to Chung-Yee Lee for his hospitality and support. The authors are also grateful to Rachel Zhang, to the senior editor, and to the referees for useful comments on an earlier version of the paper.

Table 1 **The Error in Inventory Cost due to Ignoring the Integrality of the Base-Stock Level for an Inventory Location j with Parameters r_j, $h_j = 10$, $b_j = 40$**

r_j	$\lfloor s_j^* \rfloor$	s_j^*	$s_j^* - \lfloor s_j^* \rfloor$	$(s_j^* - \lfloor s_j^* \rfloor)/s_j^*$ (%)	$f(\lfloor s_j^* \rfloor)$	$f(s_j^*)$	$f(s_j^*) - f(\lfloor s_j^* \rfloor)$	$(f(s_j^*) - f(\lfloor s_j^* \rfloor))/f(s_j^*)$ (%)
0.1	0	0.70	0.70	NA	4.44	6.99	2.55	57.27
0.2	1	1.00	0.00	0.00	10.00	10.00	0.00	0.00
0.3	1	1.34	0.34	33.68	12.14	13.37	1.22	10.09
0.4	1	1.76	0.76	75.65	16.67	17.56	0.90	5.39
0.5	2	2.32	0.32	16.10	22.50	23.22	0.72	3.20
0.6	3	3.15	0.15	5.02	31.20	31.51	0.31	0.98
0.7	4	4.51	0.51	12.81	44.68	45.12	0.45	1.00
0.8	7	7.21	0.21	3.04	71.94	72.13	0.18	0.25
0.9	15	15.28	0.28	1.84	152.65	152.76	0.10	0.07
0.95	31	31.38	0.38	1.22	313.71	313.77	0.06	0.02
0.999	1,608	1,608.63	0.63	0.04	16,086.33	16,086.33	0.00	0.00

Appendix 1. The Continuous Approximation

Let

$$f_j(\boldsymbol{\alpha}, s_j) = h_j\left[s_j - \frac{r_j}{1-r_j}\left(1 - r_j^{s_j}\right)\right] + b_j \frac{r_j^{s_j+1}}{1-r_j}$$

refer to the expected inventory cost due to location j given demand allocation $\boldsymbol{\alpha}$ and base-stock level s_j. Also, let

$$s_j^* = \frac{\ln[\omega_j]}{\ln[r_j]}.$$

Then, the error in the inventory cost due to ignoring the integrality of the base-stock level is given by

$$f_j(\boldsymbol{\alpha}, s_j^*) - f_j(\boldsymbol{\alpha}, \lfloor s_j^* \rfloor) \le f_j(\boldsymbol{\alpha}, \lfloor s_j^* \rfloor + 1) - f_j(\boldsymbol{\alpha}, \lfloor s_j^* \rfloor) \le h_j(1 - r_j) \le h_j(1 - \rho).$$

We now can verify that

$$\lim_{\rho \to 1}[f_j(\boldsymbol{\alpha}, s_j^*) - f_j(\boldsymbol{\alpha}, \lfloor s_j^* \rfloor)] = 0$$

and

$$\lim_{\rho \to 0}[f_j(\boldsymbol{\alpha}, s_j^*) - f_j(\boldsymbol{\alpha}, \lfloor s_j^* \rfloor)] \le h_j.$$

Hence, the difference in cost is largest when ρ approaches zero, but remains bounded by h_j. To put this bound in perspective, h_j corresponds to the expected cost in a system in which average inventory is equal to one and there is never a backorder. Furthermore, as illustrated in Tables 1 and 2, this bound tends to be loose, with the actual difference being significantly smaller. Note that the relative difference in cost is also bounded with

$$\frac{f_j(\boldsymbol{\alpha}, s_j^*) - f_j(\boldsymbol{\alpha}, \lfloor s_j^* \rfloor)}{f_j(\boldsymbol{\alpha}, s_j^*)} \le \frac{(1 - r_j)\ln r_j}{\ln[\omega_j]} \le \frac{(1-\rho)\ln \rho}{\ln[\omega_j]}.$$

Note that for reasonable ranges of parameters, the relative difference in cost is small. For example, for $\rho \ge 0.75$ and $\omega_j \ge 0.2$, this difference is less than 0.045. When the utilization is small, this relative difference can be large, but in that case, the absolute cost difference tends to be small. Finally, note that the bound on the relative difference decreases with

Table 2 **The Error in Inventory Cost due to Ignoring the Integrality of the Base-Stock Level for an Inventory Location j with Parameters r_j, $h_j = 10$, $b_j = 90$**

r_j	$\lfloor s_j^* \rfloor$	s_j^*	$s_j^* - \lfloor s_j^* \rfloor$	$(s_j^* - \lfloor s_j^* \rfloor)/s_j^*$ (%)	$f(\lfloor s_j^* \rfloor)$	$f(s_j^*)$	$f(s_j^*) - f(\lfloor s_j^* \rfloor)$	$(f(s_j^*) - f(\lfloor s_j^* \rfloor))/f(s_j^*)$ (%)
0.1	1	1.00	0.00	NA	10.00	10.00	0.00	0.00
0.2	1	1.43	0.43	43.07	12.50	14.31	1.81	14.45
0.3	1	1.91	0.91	91.25	18.57	19.12	0.55	2.98
0.4	2	2.51	0.51	25.65	24.00	25.13	1.13	4.71
0.5	3	3.32	0.32	10.73	32.50	33.22	0.72	2.21
0.6	4	4.51	0.51	12.69	44.44	45.08	0.64	1.43
0.7	6	6.46	0.46	7.59	64.12	64.56	0.44	0.68
0.8	10	10.32	0.32	3.19	102.95	103.19	0.24	0.23
0.9	21	21.85	0.85	4.07	218.48	218.54	0.07	0.03
0.95	44	44.89	0.89	2.02	448.88	448.91	0.03	0.01
0.999	2,301	2,301.43	0.43	0.02	23,014.33	23,014.34	0.00	0.00

the ratio $b_j/(b_j+h_j)$ and approaches zero as $b_j/(b_j+h_j)$ approaches one.

Appendix 2. Solution Procedure

Noting that $\alpha_{ij} \in \{0, 1\}$, we can restate the DAP-D as follows:

$$\text{minimize} \quad \sum_{j=1}^{m} h_j \frac{\ln[\omega_j]}{\ln[r_j]} + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\lambda_i\alpha_{ij} \tag{A1}$$

$$\text{subject to} \quad \sum_{j=1}^{m} \alpha_{ij} = 1, \quad i = 1, \ldots, n, \tag{A2}$$

$$\alpha_{ij} \in \{0, 1\}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m. \tag{A3}$$

Let $\beta_j = (\mu - \lambda)/(\sum_{i=1}^{n} \lambda_i\alpha_{ij})$ and $a_j = h_j \ln[(h_j + b_j)/b_j]$; then, we can further restate the problem as

$$\text{minimize} \quad \sum_{j=1}^{m} \frac{a_j}{\ln[1+\beta_j]} + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\lambda_i\alpha_{ij} \tag{A4}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \lambda_i\alpha_{ij}\beta_j = (\mu - \lambda)\delta_j, \quad j = 1, \ldots, m, \tag{A5}$$

$$\delta_j \in \{0, 1\}, \quad j = 1, \ldots, m, \tag{A6}$$

$$\alpha_{ij} \le \delta_j, \quad i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{A7}$$

$$\beta_j \ge 0, \quad j = 1, \ldots, m, \tag{A8}$$

$$\text{(A2), (A3),}$$

where δ_j is a binary variable that takes value one if facility j is used. Because the term $\alpha_{ij}\beta_j$ is nonlinear, we linearize it by defining $\gamma_{ij} = \alpha_{ij}\beta_j$ and substituting (A5) by

$$\sum_{i=1}^{n} \lambda_i\gamma_{ij} = (\mu - \lambda)\delta_j \quad j = 1, \ldots, m, \tag{A9}$$

$$0 \le \gamma_{ij} \le \beta_j \text{ for all } i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{A10}$$

$$\beta_j + M(\alpha_{ij} - 1) \le \gamma_{ij} \le M\alpha_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{A11}$$

where M is a large number (note that we have taken advantage of the fact that $\alpha_{ij} \in \{0, 1\}$).

Being convex, the function $f(\beta_j) = 1/\ln(1+\beta_j)$ for $j = 1, \ldots, m$, can be written as the maximum of a set of tangent piecewise linear functions. In particular, consider nonnegative points β_j^u indexed by set U. Then, $f(\beta_j) = \max_{u\in U}\{f(\beta_j^u) + f'(\beta_j)(\beta_j - \beta_j^u)\}$, where U is the index set of all possible points β_j^u. Hence, the DAP-D can be reformulated as the following optimization problem:

$$\text{minimize} \sum_{j=1}^{m} a_j \max_{u\in U}\left\{\frac{1}{\ln(1+\beta_j^u)} - \frac{\beta_j - \beta_j^u}{(1+\beta_j^u)[\ln(1+\beta_j^u)]^2}\right\} + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\lambda_i\alpha_{ij}, \tag{A12}$$

subject to constraints (A2), (A3), (A6), and (A7)–(A11). We refer to this equivalent optimization problem as the DAP-D(U). The objective function of the DAP-D(U) can be linearized by substituting the terms

$$\max_{u\in U}\left\{\frac{1}{\ln(1+\beta_j^u)} - \frac{\beta_j - \beta_j^u}{(1+\beta_j^u)[\ln(1+\beta_j^u)]^2}\right\}$$

with nonnegative decision variables θ_j and introducing the following linear constraints:

$$\theta_j \ge \frac{1}{\ln(1+\beta_j^u)} - \frac{\beta_j - \beta_j^u}{(1+\beta_j^u)[\ln(1+\beta_j^u)]^2}, \quad j = 1, \ldots, m, \text{ and all } u \in U \tag{A13}$$

$$\theta_j \ge 0, \quad j = 1, \ldots, m. \tag{A14}$$

The DAP-D(U) can now be restated as the following MILP:

$$\text{minimize} \sum_{j=1}^{m} a_j\theta_j + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\lambda_i\alpha_{ij}, \tag{A15}$$

subject to constraints (A2), (A3), (A6)–(A11), (A13), and (A14).

The nonlinearity of DAP-D was eliminated at the expense of having to deal with an exponential number of constraints. However, to solve the DAP-D(U), it is not necessary to generate all constraints (A13). Instead, it suffices to start with a subset of these constraints and generate the rest as needed. More specifically, suppose that at iteration q, $q \ge 1$, we use a subset of points indexed by U^q and solve the corresponding problem DAP-D(U^q), which yields the solution ($\boldsymbol{\alpha}^q$, $\boldsymbol{\delta}^q$, $\boldsymbol{\gamma}^q$, $\boldsymbol{\beta}^q$, $\boldsymbol{\theta}^q$) and objective function $z(U^q)$. Then, $z(U^q)$ is a lower bound on the optimal objective of the DAP-D because $z(U^q)$ is a relaxation of the DAP-D. Furthermore, $\boldsymbol{\alpha}^q$ is feasible to the DAP-D, and so

$$\sum_{j=1}^{m} a_j \frac{1}{\ln\left(1 + (\mu - \lambda)/(\sum_{i=1}^{n} \lambda_i\alpha_{ij}^q)\right)} + \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\lambda_i\alpha_{ij}$$

provides an upper bound to the DAP-D. If the upper bound (UB) and the lower bound (LB) are equal, then ($\boldsymbol{\alpha}^q$, $\boldsymbol{\delta}^q$, $\boldsymbol{\gamma}^q$, $\boldsymbol{\beta}^q$, $\boldsymbol{\theta}^q$) is an optimal solution for the DAP-D(U) and also for the DAP-D. Otherwise, a new set of constraints (A13) is generated at

$$\beta_j^{\text{new}} = \begin{cases} \dfrac{\mu - \lambda}{\sum_{i=1}^{n} \lambda_i\alpha_{ij}^q} & \text{if } \delta_j^q = 1 \\ M & \text{if } \delta_j^q = 0, \end{cases}$$

a new problem DAP-D(U^{q+1}) with the new constraints added to the current set is generated. The procedure is repeated until the lower and upper bounds coincide (i.e., UB = LB). Note that although the lower bound is monotonic, the upper bound is not, and so the best upper bound needs to be stored. It is not difficult to show that at each iteration, at least one new point and corresponding constraints

Table 3 Numerical Results for Problems Without Fixed Costs

			$\tau = 10$			$\tau = 20$		
n	m	ρ	CPU time (seconds)	Number of iterations	Optimal cost	CPU time (seconds)	Number of iterations	Optimal cost
30	10	0.6	0.41	2	327.46	0.73	2	171.90
		0.7	0.50	2	331.29	0.77	2	175.73
		0.8	0.56	2	338.31	25.71	6	182.75
		0.9	1.27	2	357.75	>1,000 (1.79%)	3	202.96
30	15	0.6	8.07	4	215.91	8.77	3	115.50
		0.7	13.42	5	219.62	26.91	5	119.21
		0.8	17.43	4	226.44	25.54	4	126.03
		0.9	32.54	4	245.49	>1,000 (4.1%)	3	147.98
30	20	0.6	0.97	2	110.93	1.03	2	68.50
		0.7	0.92	2	115.65	1.63	2	73.22
		0.8	1.11	2	123.83	994.90	6	81.41
		0.9	2.15	2	145.06	>1,000 (8.63%)	3	104.39
40	10	0.6	0.51	2	694.76	0.84	2	355.35
		0.7	0.50	2	698.56	1.35	2	359.14
		0.8	1.08	2	705.54	6.14	2	366.13
		0.9	1.58	2	724.98	>1,000 (4.47%)	2	399.56
40	15	0.6	5.31	3	512.6	7.29	3	263.70
		0.7	8.10	3	516.29	34.89	3	267.38
		0.8	11.61	3	523.10	32.04	4	274.20
		0.9	41.37	4	542.17	>1,000 (1.21%)	3	294.44
40	20	0.6	1.30	2	336.38	1.29	2	180.81
		0.7	1.31	2	341.01	1.77	2	185.45
		0.8	1.21	2	349.11	25.74	2	193.55
		0.9	2.94	2	370.31	>1,000 (5.4%)	3	217.13
50	10	0.6	0.51	2	1,203.59	1.19	2	609.62
		0.7	0.67	2	1,207.35	2.33	2	613.38
		0.8	1.57	2	1,214.30	3.12	2	620.33
		0.9	1.69	2	1,233.72	125.21	7	639.75
50	15	0.6	5.93	3	950.77	7.43	3	482.67
		0.7	12.01	4	954.44	18.17	4	486.33
		0.8	13.12	3	961.24	30.37	3	493.14
		0.9	63.67	4	980.33	600.54	5	512.22
50	20	0.6	1.19	2	703.43	1.68	2	364.02
		0.7	1.65	2	707.99	1.70	2	368.58
		0.8	1.60	2	716.02	3.53	2	376.61
		0.9	3.27	2	737.17	>1,000 (1.28%)	4	400.71

Notes. For cases where the CPU time exceeds 1,000 seconds, we report the cost of the best solution and the maximum percentage error $= 100\%(\text{UB} - \text{LB})/\text{LB}$. $c_{ij} = |2i - 2j + m - n|/(\sqrt{2}\tau)\ \forall i, j$, $\lambda_i = 20\ \forall i$, $h_j = 2$, $b_j = 10$, $\forall j$.

are introduced. Therefore, cycling cannot occur. Moreover, because the variables α_{ij} are binary, there is a finite set of values that β_j can take. Consequently, an optimal solution is always reached within a finite number of iterations.

The above approach can be adapted immediately to solve problems with fixed location costs discussed in §5 (for brevity, we omit repeating the details). In Tables 3 and 4, we provide representative numerical results illustrating the computational effectiveness of the solution procedure for systems with and without fixed location costs. The procedure was coded in Matlab 7.0 with the solution to the MILP problems obtained using CPLEX 10.1 on a Sun Blade 2500 workstation.

It is clear from Tables 3 and 4 that as ρ or τ increases, the problem becomes more difficult to solve. A possible explanation is that as either ρ or τ increases, the transportation cost component in the objective function becomes less dominant, requiring more search effort for an optimal solution (recall from the discussion in §1 that when transportation costs are dominant, the problem reduces to a simple assignment problem that is easy to solve). Problems with fixed location costs appear to be relatively easier to solve than

Table 4 Numerical Results for Problems with Fixed Costs

			$\tau = 10$			$\tau = 20$		
n	m	ρ	CPU time (seconds)	Number of iterations	Optimal cost	CPU time (seconds)	Number of iterations	Optimal cost
30	10	0.6	2.34	5	409.10	0.68	2	232.31
		0.7	2.37	5	412.33	1.25	2	235.43
		0.8	5.29	7	418.54	2.58	3	241.52
		0.9	13.98	7	436.72	22.58	4	259.56
30	15	0.6	10.43	8	320.02	11.55	7	195.62
		0.7	13.17	7	323.34	12.91	7	198.84
		0.8	45.63	10	329.68	91.03	8	205.05
		0.9	101.01	12	348.00	222.66	12	223.21
30	20	0.6	35.40	12	259.25	242.10	14	175.85
		0.7	40.29	12	262.68	333.26	13	179.07
		0.8	146.06	12	269.13	414.63	15	185.27
		0.9	201.85	13	287.60	680.25	17	203.43
40	10	0.6	2.47	5	776.75	1.40	2	416.15
		0.7	2.96	5	779.97	1.54	2	419.27
		0.8	5.82	7	786.19	4.15	3	425.37
		0.9	16.19	7	804.38	16.54	3	443.41
40	15	0.6	9.95	7	616.95	4.96	5	344.07
		0.7	9.59	7	620.23	15.80	7	347.29
		0.8	36.37	11	626.57	36.86	8	353.50
		0.9	100.04	11	644.92	96.21	9	371.68
40	20	0.6	39.30	12	485.42	56.85	10	288.93
		0.7	41.19	11	488.84	138.86	11	292.19
		0.8	150.16	13	495.30	179.41	10	298.40
		0.9	229.52	12	513.79	566.54	13	316.56
50	10	0.6	2.68	5	1,285.83	1.15	2	670.71
		0.7	4.39	5	1,289.05	1.11	2	673.83
		0.8	5.62	5	1,295.26	2.05	2	679.93
		0.9	16.06	7	1,313.46	21.30	3	697.97
50	15	0.6	11.03	8	1,055.24	8.22	5	563.26
		0.7	12.32	7	1,058.60	21.01	6	566.46
		0.8	45.02	8	1,064.90	45.95	8	572.67
		0.9	124.93	11	1,083.26	183.39	13	590.86
50	20	0.6	43.67	12	853.01	172.26	12	472.72
		0.7	50.67	13	856.43	144.71	11	476.01
		0.8	87.74	12	862.89	214.52	11	482.22
		0.9	200.89	12	881.40	434.62	11	500.39

Note. $c_{ij} = |2i - 2j + m - n|/(\sqrt{2}\tau)\ \forall i, j$, $\lambda_i = 20\ \forall i$, $h_j = 2$, $b_j = 10$, $K_j = 20$, $\forall j$.

those without such costs. This may be because adding such costs diminishes the importance of inventory costs.

References

Albin, S. L. 1984. Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Oper. Res.* **32** 1133–1162.

Bassok, Y., R. Anupindi, R. Akella. 1999. Single-period multiproduct inventory models with substitution. *Oper. Res.* **47** 632–642.

Bell, C. E., C. Stidham. 1983. Individual versus social optimization in the allocation of customers to alternative servers. *Management Sci.* **29** 831–839.

Benjaafar, S., D. Gupta. 1999. Workload allocation in multi-product, multi-facility production systems with setup times. *IIE Trans.* **31** 339–352.

Benjaafar, S., W. L. Cooper, J. S. Kim. 2005. On the benefits of inventory pooling in production-inventory systems. *Management Sci.* **51** 548–565.

Benjaafar, S., M. ElHafsi, F. de Véricourt. 2004. Demand allocation in multiple-product, multiple-facility make-to-stock systems. *Management Sci.* **50** 1431–1448.

Bonomi, F., A. Kumar. 1990. Adaptive optimal load balancing in a nonhomogeneous multiserver system with a central job scheduler. *IEEE Trans. Comput.* **39** 1232–1250.

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

Cornuéjols, G., G. L. Nemhauser, L. A. Wolsey. 1990. The uncapacitated facility location problem. P. Mirchandani, R. Francis, eds. *Discrete Location Theory*. John Wiley & Sons, New York, 119–171.

Daskin, M. S., L. V. Snyder, R. T. Berger. 2005. Facility location in supply chain design. A. Langevin, D. Riopel, eds. *Logistics Systems: Design and Operation*. Springer, New York, 39–66.

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

Eppen, G. D. 1979. Effects of centralization on expected costs in a multi-location newsboy problem. *Management Sci.* **25** 498–501.

Gerchak, Y., Q.-M. He. 2003. On the relation between the benefits of risk pooling and the variability of demand. *IIE Trans.* **35** 1027–1031.

Gerchak, Y., M. Henig. 1989. Component commonality in assemble-to-order system: Models and properties. *Naval Res. Logist.* **36** 61–68.

Labbé, M., F. V. Louveaux. 1997. Location problems. M. Dell'Amico, F. Maffioli, S. Martello, eds. *Annotated Bibliographies in Combinatorial Optimization*. Wiley, Chichester, UK, 261–281.

Liu, Z., R. Righter. 1998. Optimal load balancing on distributed homogeneous unreliable processors. *Oper. Res.* **46** 563–573.

Murota, K. 2003. *Discrete Convex Analysis* (*SIAM Monographs Discrete Math. Appl.*). Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Netessine, S., G. Dobson, R. A. Shumsky. 2002. Flexible service capacity: Optimal investment and the impact of demand correlation. *Oper. Res.* **50** 375–388.

Ni, L. M., K. Hwang. 1985. Optimal load balancing in multiple processor system with many job classes. *IEEE Trans. Software Engrg.* **SE-11** 491–496.

Peterson, W. P. 1991. A heavy traffic limit theorem for networks of queues with multiple customer types. *Math. Oper. Res.* **16** 90–118.

Shen, Z.-J. M., C. R. Coullard, M. S. Daskin. 2003. A joint location-inventory model. *Transportation Sci.* **37** 40–55.

Sherali, H. D., I. Al-Loughani, S. Subramanian. 2002. Global optimization procedures for the capacitated Euclidean and l_p distance multifacility location-allocation problems. *Oper. Res.* **50** 433–448.

Tang, C. S., M. van Vliet. 1994. Traffic allocation for manufacturing systems. *Eur. J. Oper. Res.* **75** 171–185.

Thonemann, U. W., M. L. Brandeau. 2000. Optimal commonality in component design. *Oper. Res.* **48** 1–19.

Tijms, H. 1995. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, New York.

van Houtum, G., I. Adan, J. van der Wal. 1997. The symmetric longest queue system. *Stochastic Models* **13** 105–120.

van Mieghem, J. A., N. Rudi. 2002. Newsvendor networks: Inventory management and capacity investment with discretionary activities. *Manufacturing Service Oper. Management* **4** 313–335.

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

Wang, Y., R. J. T. Morris. 1985. Load sharing in distributed systems. *IEEE Trans. Comput.* **34** 204–217.

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

Whitt, W. 1982. Approximating a point process by a renewal process, I: Two basic approaches. *Oper. Res.* **30** 125–147.

Whitt, W. 1983. The queueing network analyzer. *Bell System Tech. J.* **62** 2779–2815.

Whitt, W. 1993. Approximations for the GI/G/m queue. *Production Oper. Management* **2** 114–161.

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

Zheng, Y.-S., P. Zipkin. 1990. A queueing model to analyze the value of centralized inventory information. *Oper. Res.* **38** 296–307.

Zipkin, P. 1995. Performance analysis of a multi-item production-inventory system under alternative policies. *Management Sci.* **41** 690–703.

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