MANAGEMENT SCIENCE
Vol. 50, No. 10, October 2004, pp. 1431–1448
ISSN 0025-1909 | EISSN 1526-5501 | 04 | 5010 | 1431

informs

DOI 10.1287/mnsc.1040.0250

Demand Allocation in Multiple-Product, Multiple-Facility, Make-to-Stock Systems

Saif Benjaafar
Graduate Program in Industrial Engineering, Department of Mechanical Engineering, University of Minnesota, Minneapolis, Minnesota 55455-0111, saif@umn.edu

Mohsen ElHafsi
The A. Gary Anderson Graduate School of Management, University of California, Riverside, California 92521-0203, mohsen.elhafsi@ucr.edu

Francis de Véricourt
The Fuqua School of Business, Duke University, Durham, North Carolina 27708, fdv1@duke.edu

We consider the problem of allocating demand arising from multiple products to multiple production facilities with finite capacity and load-dependent lead times. Production facilities can choose to manufacture items either to stock or to order. Products vary in their demand rates, holding and backordering costs, and service-level requirements. We develop models and solution procedures to determine the optimal allocation of demand to facilities and the optimal inventory level for products at each facility. We consider two types of demand allocation, one in which we allow the demand for a product to be split among multiple facilities and the other in which demand from each product must be entirely satisfied by a single facility. We also consider two forms of inventory warehousing, one in which inventory locations are factory based and one in which they are centralized. For each case, we offer a solution procedure to obtain optimal demand allocations and optimal inventory base-stock levels. For systems with multiple customer classes, we also determine optimal inventory rationing levels for each class for each product. We use the models to characterize analytically several properties of the optimal solution. In particular, we highlight eight principles that relate the effects of cost, congestion, inventory pooling, multiple sourcing, customer segmentation, inventory rationing, and process and demand variability.

Key words: production/inventory systems; queueing systems; generalized assignment problem; multi-item/multifacility systems
History: Accepted by William S. Lovejoy, operations and supply chain management; received April 12, 2002. This paper was with the authors 11 months for 4 revisions.

1. Introduction

We consider the problem of allocating demand that arises from N classes of items, say N products, to a set of M production facilities. The production facilities vary in their capacities and in their production and inventory-handling cost structures. Products vary in their demand rates, holding and backordering costs, and service-level requirements. The production facilities can either make products to stock or make them to order. Demand for each product occurs over time with stochastic interarrival times. Production times are also stochastic. Consequently, production lead times (and consequently inventory replenishment lead times) are stochastic and affected by congestion at the production facilities. Demand for the same product can originate from different customer classes. Customer classes may vary in their demand rates, backorder penalties, or service-level requirements.

We present models and solution procedures to jointly determine the optimal allocation of demand to facilities and the optimal inventory level for each product at each facility. We consider two types of demand allocation, one in which we allow the demand for a product to be split among multiple facilities, and the other in which demand from each product must be entirely satisfied by a single facility. We refer to the former as the *demand allocation problem* and the latter as the *demand partitioning problem*. Also, we consider two forms of inventory warehousing, one in which inventory locations are factory based and one in which they are centralized. For each case, we offer a solution procedure to obtain optimal demand allocations and optimal inventory base-stock levels. In cases where the same product is demanded by multiple customer classes, we also determine an optimal allocation of customer classes to facilities and an optimal rationing policy of inventory if demands from

multiple customer classes are satisfied from the same inventory buffer.

The joint inventory control and demand allocation problem with multiple products and multiple facilities is complex and involves the balancing of several trade-offs. For example, although it may appear desirable to assign demand to facilities with the lowest production and transportation costs, excessive workloads on these facilities lead to long and highly variable replenishment lead times, which in turn may require higher inventory levels or lead to higher backorder costs. A degree of workload balancing among facilities is thus necessary. This workload balancing must, however, be carried out carefully, because there can be value in centralizing inventory in as few locations as possible. Unbalanced workloads may be tolerated if they yield sufficient savings from the associated inventory pooling. Inventory pooling, on the other hand, is not always desirable. In systems where the demand for each product originates from multiple customer classes, it may be more beneficial to segment customer classes based on their backorder costs and to serve them from different facilities. Alternatively, we may ration the inventory at each location among the different classes. In this case, it may indeed be desirable to serve different customer classes from the same location. In systems where demand is highly variable, allocating the demand of each product to multiple facilities can reduce the demand variability experienced by these facilities, hence reducing their replenishment lead times. In settings where inventory for each product must be held in a centralized location, distributing demand from each product among multiple facilities can also be desirable because it reduces the variance of replenishment lead times.

One of the objectives of this paper is to highlight the effect of these various factors on demand allocation and inventory control decisions and to derive useful managerial insights and guidelines. In particular, we identify eight principles that relate the effects of cost, congestion, inventory pooling, multiple sourcing, customer segmentation, inventory rationing, and process and demand variability. We show how these principles can be used to guide demand allocation decisions even when the models themselves are not used. Furthermore, we illustrate several properties of the optimal solution, including a few that are seemingly counterintuitive.

The problem of demand allocation arises in a variety of contexts when there are multiple facilities available to manufacture the same product. In this paper, we are in part motivated by our work with a large contract manufacturer (CM) in the electronics industry. The CM, which has over 35 manufacturing facilities worldwide, offers manufacturing outsourcing services to several leading original equipment manufacturers (OEMs) for a variety of industrial and consumer products. The core manufacturing technology in most facilities is surface mount technology (SMT) for printed circuit-board assembly (PCA). To meet the needs of a wide range of OEMs, the CM has invested in flexible technologies that can accommodate a high variety of products that are manufactured in varying quantities with little changeover time or cost. Because the plants tend to vary in size, efficiency, and in geographic location, the costs of production, warehousing, and transportation also vary. Although the availability of a large number of facilities with similar manufacturing capabilities provides the CM with a high degree of flexibility, it also poses challenges as to how products should be assigned to facilities and how much inventory to keep for each product at each facility. Because the manufacturing and inventory functions are tightly coupled (many of the OEMs tend to prefer that inventory is held and managed by the CM with deliveries made on a just-in-time basis), demand-assignment models that focus on production costs do not adequately account for inventory accumulation and related costs. Similarly, commonly used inventory models where supply lead times are assumed exogenous do not fully account for the impact of factory congestion on inventory replenishment lead times.

2. Literature Review

The demand allocation problem has been widely studied in the literature in the context of make-to-order systems where the objective is typically to optimize a function of manufacturing lead time. Examples from this literature include Benjaafar and Gupta (1999), Green and Guha (1995), Tang and Van Vliet (1994), and the references therein. Several important cases are discussed in Buzacott and Shanthikumar (1993). A closely related problem is the *load-sharing problem* that arises in the design of distributed computer systems. A review of the extensive literature on this problem can be found in Wang and Morris (1985) and more recent developments are described in Liu and Righter (1998). The problem we treat in this paper generalizes the load-sharing problem to make-to-stock systems.

A common assumption in the above literature is that demand streams can be split among the available facilities; however, situations do arise where the demand for items cannot be split and the demand from each item must be assigned in its entirety to a single facility. When such restrictions apply, the problem becomes combinatorial and the decision variables discrete. The literature on this version of the problem is more limited as it is generally more difficult to solve. However, a well-studied version of the problem

is the well-known generalized assignment problem (GAP). For the GAP, all parameters are deterministic and the objective is to minimize a linear function of assignment costs subject to capacity constraints. Although the problem is NP-hard, several relatively efficient algorithms have been proposed. The reader is referred to the early work of Ross and Soland (1975), Fisher (1981), and Martello and Toth (1980). Recent reviews of GAP applications and solution algorithms can be found in Cattrysse and Van Wassenhove (1992) and Osman (1995). As we show in §4, the GAP is a special instance of the problem we model.

Our work is also related to the growing body of literature on make-to-stock queues. Example papers that deal with multiple-class systems include Wein (1992), Ha (1997), Zipkin (1995), de Véricourt et al. (2000), de Véricourt et al. (2002), and Benjaafar et al. (2004a). An excellent summary of important results is provided in Buzacott and Shanthikumar (1993). This literature, which usually treats a single facility, is mostly concerned with characterizing optimal production and inventory control policies. To our knowledge, our paper is the first to consider systems with multiple facilities and multiple products. It appears also to be the first paper to consider the joint demand allocation and inventory control problem in a setting of make-to-stock queues.

The remainder of this paper is organized as follows. In §3, we present our basic model and formulate both the demand allocation and demand partitioning problems. For the demand allocation problem, we consider two cases, one where inventory is held at the production facilities and one where it is centralized in a single location. In §4, we offer solution methods to both problems and in §5 we present numerical results. In §6, we treat several special cases and offer additional insights into the nature of the optimal solution. In §7, we extend our basic model to include multiple customer classes, heterogeneous processing times, and demand processes with general distributions. Finally, in §8, we offer concluding comments.

3. Model Description

We consider a system with M manufacturing facilities and N products. Demand for each product occurs one unit at a time according to an independent Poisson process with rate λ_i for product i ($i = 1, \ldots, N$). We let α_{ij} ($0 \le \alpha_{ij} \le 1$) denote the long-run fraction of demand of product i that is supplied by facility j. To ensure that the demand from each product is met, we require that $\sum_{j=1}^{M} \alpha_{ij} = 1$. The total demand satisfied by facility j is $\hat{\lambda}_j = \sum_{i=1}^{N} \alpha_{ij}\lambda_i$. Naturally, we have $\sum_{j=1}^{M} \hat{\lambda}_j = \sum_{i=1}^{N} \lambda_i$. In systems where demand cannot be split among multiple facilities the parameter α_{ij} is restricted to be either 0 or 1. Each facility j has a finite production capacity with rate μ_j. Unit processing times at each facility are assumed to be i.i.d. and exponentially distributed random variables with mean $1/\mu_j$. The case of product-dependent service times is discussed in §7. Demand for each product is satisfied from finished-goods inventory. If none is available, demand is backordered. First, we consider the case where inventory for each product is held at the facility that produces it (the case where inventory for each product is stocked in a single centralized location is discussed later in this section). Finished-goods inventories are managed according to a base-stock policy, with base-stock level s_{ij} for product i at facility j. Each facility j incurs a production cost c_{ij} per unit of product i (production costs may include both manufacturing and transportation costs), a holding cost h_{ij} per unit of inventory of product i per unit time, and a backordering cost b_{ij} per unit of product i backordered per unit time. To ensure overall feasibility, we assume that $\sum_{i=1}^{N} \lambda_i < \sum_{j=1}^{M} \mu_j$. To guarantee feasibility at each facility, we require that the utilization $\rho_j = \sum_{i=1}^{N} \alpha_{ij}\lambda_i/\mu_j$ of each facility j satisfies the stability condition $\rho_j < 1$. Note that because the probabilistic decomposition of a Poisson process is a Poisson process, the demand from a product i that is assigned to facility j also follows a Poisson process. Furthermore, because the superposition of Poisson processes is a Poisson process, the aggregated demand arrival process at each production facility follows also a Poisson process. Hence, viewed in isolation, each production facility j can be modeled as an $M/M/1$ queue with arrival rate $\hat{\lambda}_j$ and service rate μ_j.

Some of the above assumptions are relaxed later. In this section we consider systems with centralized inventory warehousing. In §7, we discuss extensions of this basic model to systems where demand for the same product might arise from multiple customer classes with nonidentical backorder costs or service levels and to systems with heterogeneous processing requirements and non-Markovian demands. Although it is possible to consider systems where a changeover time or cost is incurred when a production facility switches from one product to another (see, for example, Benjaafar et al. 2004b), we focus here on systems with flexible manufacturing technologies where setups are relatively insignificant (in this sense, our setup is similar to that of Zipkin 1995 and Wein 1992).

Throughout this paper we assume that orders are processed at the production facilities on a first-come–first-served (FCFS) basis. Our use of the FCFS policy is motivated by its widespread use in practice, its ease of implementation, its perceived fairness, and its analytical tractability. Characterizing an optimal policy for a production-inventory system with multiple products is a difficult problem that to date remains

unresolved for the general case; see de Véricourt et al. (2000) for results and references. In principle, the problem can be formulated as a Markov decision process (MDP) and solved numerically. However, because of the well-known curse of dimensionality for multidimensional MDPs, an optimal policy cannot be identified in any reasonable amount of time, except for the smallest systems (e.g., a single facility with two products) and for relatively low utilization levels. Nevertheless, there is evidence that the difference in cost between the FCFS and an optimal policy diminishes in utilization, with this difference becoming negligible when utilization is high (see Wein 1992, Zheng and Zipkin 1990, Zipkin 1995, or Van Houtum et al. 1997). Static priorities among the different products can also provide an alternative to the FCFS policy. However, static priorities are analytically difficult to evaluate in a make-to-stock setting, and can sometimes be less efficient than the FCFS policy; see de Véricourt and Dallery (2000) and de Véricourt and Veatch (2003).

Our objective is to identify an allocation matrix $\boldsymbol{\alpha}^* = \{\alpha_{ij}^*\}$ and base-stock level matrix $\mathbf{s}^* = \{s_{ij}^*\}$ so that the long-run expected total cost per unit time is minimized. We denote this expected total cost by

$$z(\boldsymbol{\alpha}, \mathbf{s}) = \sum_{i=1}^{N}\sum_{j=1}^{M}\{h_{ij}E(I_{ij}) + b_{ij}E(B_{ij}) + c_{ij}\alpha_{ij}\lambda_i\}, \qquad (1)$$

where I_{ij} and B_{ij} are random variables equal in distribution to, respectively, the steady-state inventory and backorder levels for product i at facility j. Instead of minimizing the sum of holding and backordering costs, inventory and demand allocation policies may be chosen so that only the inventory-holding and production cost components of the total expected cost is minimized with a requirement that a specified service level for each product at each facility is met. A service level can be specified in several ways, including setting an upper bound on the probability of a stockout or choosing a target fill rate (the fraction of orders filled from stock). In our case, because demand occurs one unit at a time, the requirement on either the stockout probability or the fill rate can be specified as a constraint on the probability that an arriving order of type i finds no inventory on hand. This is also equivalent to finding s_{ij} units already on order. If we let Q_{ij} denote the number of units on order of type i at facility j as seen by an arrival, then the service-level constraint can be stated as

$$\Pr(Q_{ij} \geq s_{ij}) \leq \gamma_{ij}, \qquad (2)$$

where $\gamma_{ij} \in [0, 1]$ is the specified service level for product i at facility j.

We consider two allocation policies: *probabilistic* and *deterministic*. Under the former, the variable α_{ij} is allowed to assume values over the range [0, 1]. In this case, α_{ij} also corresponds to the probability that incoming demand of type i is assigned to facility j. Although a truly probabilistic allocation is unlikely in practice, it is useful in approximating the behavior of a central dispatcher that attempts to adhere to a specified workload for each facility. It is also useful in modeling settings where demand for each product arises from a sufficiently large number of sources. The variable α_{ij} corresponds in that case to the fraction of demand sources (e.g., geographical locations) for type i that is always satisfied from facility j. Under a deterministic policy, we restrict the variables α_{ij}s to be either 0 or 1 so that each product is produced in only one facility. However, a facility may produce multiple products under both policies. In the remainder of this article, we shall refer to the problem with probabilistic allocation as the demand allocation problem (DAP) and to the one with deterministic allocation as the demand partitioning problem (DPP).

In Lemma 1, we show how to obtain expressions for expected inventory, expected backorder level, and probability of backordering (or fill rate) for a given base-stock level matrix $\mathbf{s}$ and allocation matrix $\boldsymbol{\alpha}$. The proof is in Appendix 1.

Lemma 1. *Given* $\boldsymbol{\alpha} = \{\alpha_{ij}\}$ *and* $\mathbf{s} = \{s_{ij}\}$, *expected inventory level, number of backorders, and probability of backordering under the DAP assumptions are given, respectively, by* $E(I_{ij}) = s_{ij} - (1 - r_{ij}^{s_{ij}})r_{ij}/(1 - r_{ij})$, $E(B_{ij}) = r_{ij}^{s_{ij}+1}/(1 - r_{ij})$, *and* $\Pr(Q_{ij} \geq s_{ij}) = r_{ij}^{s_{ij}}$ *where* $r_{ij} = \alpha_{ij}\lambda_i/(\mu_j - \sum_{k\neq i}\alpha_{kj}\lambda_k)$.

The DAP can now be formulated as follows:

DAP: minimize $z(\boldsymbol{\alpha}, \mathbf{s})$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{h_{ij}\left[s_{ij} - \left(\frac{r_{ij}}{1 - r_{ij}}\right)(1 - r_{ij}^{s_{ij}})\right] + b_{ij}\left[\frac{r_{ij}^{s_{ij}+1}}{1 - r_{ij}}\right] + c_{ij}\alpha_{ij}\lambda_i\right\} \qquad (3)$$

subject to

$$\sum_{j=1}^{M}\alpha_{ij} = 1, \quad i = 1, 2, \ldots, N, \qquad (4)$$

$$\sum_{i=1}^{N}\{\alpha_{ij}\lambda_i/\mu_j\} - 1 \leq 0, \quad j = 1, 2, \ldots, M, \qquad (5)$$

$$\alpha_{ij} \geq 0, \quad i = 1, 2, \ldots, N, \; j = 1, 2, \ldots, M, \qquad (6)$$

s_{ij}: positive integer,

$$i = 1, 2, \ldots, N, \; j = 1, 2, \ldots, M. \qquad (7)$$

Constraint (4) ensures that the demand for each product is met and constraint (5) ensures that capacity

constraints are not violated. For the DPP, constraint (6) is replaced by

$$\alpha_{ij} = 0, 1, \quad i = 1, 2, \ldots, N, \ j = 1, 2, \ldots, M. \tag{8}$$

For systems where a service-level constraint is in effect, the backordering cost component is eliminated from the objective function and a constraint in the form of (2) is introduced for each product at each facility.

Given an allocation matrix $\boldsymbol{\alpha}$, it is not too difficult to show that $z(\boldsymbol{\alpha}, \mathbf{s})$ is jointly convex in the s_{ij}s. Noting that z is also separable in the s_{ij}s, the optimal base-stock level s_{ij}^* for each product i at each facility j can be obtained as the smallest integer that satisfies the constraint $z_{ij}(\boldsymbol{\alpha}, s_{ij}+1) - z_{ij}(\boldsymbol{\alpha}, s_{ij}) \geq 0$ where $z_{ij}(\boldsymbol{\alpha}, s_{ij})$ is the cost contribution of product i at facility j. This leads to

$$s_{ij}^* = \lfloor \tilde{s}_{ij} \rfloor \quad \text{where } \tilde{s}_{ij} = \frac{\ln[h_{ij}/(h_{ij}+b_{ij})]}{\ln[r_{ij}]}. \tag{9}$$

In the remainder of this paper we relax the integrality on s_{ij}^* and let $s_{ij}^* = \tilde{s}_{ij}$, which simplifies the analysis considerably. The amount of error introduced by this relaxation is negligible when the s_{ij}^*s are large, which would be the case when the utilization of the production facility is high (note that $r_{ij} \to 1$ as $\rho_j \to 1$). In practice, when the utilization is low, the number of facilities is usually small because it is uncommon to maintain a large number of underutilized facilities. Consequently, the optimization problem is trivial in that case and can be solved by exhaustive search. Furthermore, 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 Shanthikumar 1993, Wein 1992, Zipkin 1995).

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

$$\begin{aligned} z(\boldsymbol{\alpha}, \mathbf{s}^*) &= \sum_{i=1}^{N}\sum_{j=1}^{M}\{h_{ij}s_{ij}^* + c_{ij}\alpha_{ij}\lambda_i\} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\frac{h_{ij}\ln[h_{ij}/(h_{ij}+b_{ij})]}{\ln[r_{ij}]} + c_{ij}\alpha_{ij}\lambda_i\right\}. \end{aligned} \tag{10}$$

When a service-level constraint is enforced, we minimize

$$\hat{z}(\boldsymbol{\alpha}, \mathbf{s}) = \sum_{i=1}^{N}\sum_{i=1}^{M}\left\{h_{ij}[s_{ij} - (1-r_{ij}^{s_{ij}})r_{ij}/(1-r_{ij})] + c_{ij}\alpha_{ij}\lambda_{ij}\right\}, \tag{11}$$

subject to (2) and (4)–(6). Rewriting constraint (2) as $s_{ij} \geq \ln(\gamma_{ij})/\ln(r_{ij})$ and noting that the objective function is strictly increasing (and convex) in the s_{ij}s, we can see that (2) is binding and the optimal base-stock level is given by $s_{ij}^* = \ln(\gamma_{ij})/\ln(r_{ij})$. Substituting s_{ij}^* in the objective function, we can now express the optimal cost for a given assignment matrix $\boldsymbol{\alpha}$ as follows:

$$\hat{z}(\boldsymbol{\alpha}, \mathbf{s}^*) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{h_{ij}\left(\frac{\ln(\gamma_{ij})}{\ln(r_{ij})} - \frac{r_{ij}(1-\gamma_{ij})}{1-r_{ij}}\right) + c_{ij}\alpha_{ij}\lambda_i\right\}. \tag{12}$$

Rewriting the objective function in the form of either (10) or (12) allows us to reduce our optimization problem to that of finding $\boldsymbol{\alpha}^*$ subject to constraints (4)–(6) (or (4), (5), and (8) in the DPP case).

In settings where finished-goods inventory is kept only in centralized warehouses and not at the factories, there is only one inventory location per product. This inventory is replenished by multiple facilities if the demand for that product gets allocated to more than one facility. In other words, inventory is always pooled while production can be distributed. We refer to this problem as the demand allocation problem with centralized inventory, or DAP-C. We let α_{ij} denote the fraction of demand of product i assigned to facility j (as in the DAP, α_{ij} can be viewed as the probability with which a demand for product i is replenished from facility j) and s_i denote the base-stock level for product i. Then, as in the DAP, each production facility j can be viewed as an independent $M/M/1$ queue with arrival rate $\sum_{i=1}^{N}\alpha_{ij}\lambda_i$. We use h_i and b_i to refer, respectively, to the holding and backorder costs per unit, per unit time, for product i.

In Lemma 2, we show how to obtain expressions for expected inventory, expected backorder level, and probability of backordering for a given base-stock level vector $\mathbf{s}$ and allocation matrix $\boldsymbol{\alpha}$. The proof is in Appendix 2.

LEMMA 2. *Given* $\boldsymbol{\alpha} = \{\alpha_{ij}\}$ *and* $\mathbf{s} = \{s_i\}$, *expected inventory level, number of backorders, and probability of backordering under the DAP-C assumptions are given, respectively, by* $E(I_i) = \sum_{j=1}^{M} R_{ij}[s_i - r_{ij}(1 - r_{ij}^{s_i})/(1-r_{ij})]$, $E(B_i) = \sum_{j=1}^{M} R_{ij}[r_{ij}^{s_i+1}/(1-r_{ij})]$, *and* $\Pr(Q_i \geq s_i) = \sum_{j=1}^{M} R_{ij} r_{ij}^{s_i}$, *where* $R_{ij} = r_{ij}^{M-1}\prod_{k\neq j}(1-r_{ik})/(r_{ij}-r_{ik})$.

The DAP-C can now be formulated as

$$\begin{aligned} \text{DAP-C:} \quad \min \ z(\boldsymbol{\alpha}, \mathbf{s}) = &\sum_{i=1}^{N} h_i\left[s_i - \sum_{j=1}^{M} R_{ij}\frac{r_{ij}}{(1-r_{ij})}\right] \\ &+ (h_i + b_i)\sum_{j=1}^{M} R_{ij}\frac{r_{ij}^{s_i+1}}{(1-r_{ij})} \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\alpha_{ij}\lambda_i, \end{aligned} \tag{13}$$

subject to constraints (4)–(6) and s_i: positive integer for $i = 1, 2, \ldots, N$.

Given an allocation matrix $\boldsymbol{\alpha}$, the optimal base-stock level s_i^* for product i can be obtained as the smallest integer s_i that satisfies $z_i(\boldsymbol{\alpha}, s_i + 1) - z_i(\boldsymbol{\alpha}, s_i) \geq 0$, or equivalently,

$$\sum_{j=1}^{M} R_{ij} r_{ij}^{s_i} \leq \frac{h_i}{h_i + b_i}. \quad (14)$$

Although a closed-form expression for s_i^* is difficult to obtain, it is straightforward to compute numerically from (14) for a given allocation matrix $\boldsymbol{\alpha}$.

Clearly, with regard to production- and inventory-related costs, a system with centralized warehousing is superior to one where inventory is factory based (a sample path argument can be easily constructed to support this claim). Whether this configuration is ultimately more desirable would of course depend on the economics of centralized versus factory-based warehousing.

4. Solution Procedure

In this section, we present procedures for solving the DAP, the DPP, and the DAP-C. For the DAP, the total cost function in (10) is not jointly convex in the decision variables α_{ij}. Hence, the DAP is difficult to solve directly. In what follows, we show how the problem can be solved via a change of variables that transforms it into a problem that is jointly convex in its decision variables. In particular, we introduce the new variables y_{ij} and rewrite the original problem as follows:

minimize

$$z(\boldsymbol{\alpha}, \mathbf{y}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left\{ c_{ij} \alpha_{ij} \lambda_i - \frac{h_{ij} \ln(h_{ij}/h_{ij} + b_{ij})}{y_{ij}} \right\}, \quad (15)$$

subject to (4), (5), and (6),

$$\lambda_i \alpha_{ij} e^{y_{ij}} + \sum_{k \neq i} \alpha_{kj} \lambda_k - \mu_j \leq 0$$

$$i = 1, 2, \ldots, N, \ j = 1, 2, \ldots, M; \quad \text{and} \quad (16)$$

$$y_{ij} \geq 0, \quad i = 1, 2, \ldots, N, \ j = 1, 2, \ldots, M. \quad (17)$$

First, note that constraints (16) are always binding at an optimal solution because the y_{ij} variables are always chosen to be as large as possible. Because the α_{ij}s and y_{ij}s are independently expressed in the objective function, the objective function is jointly convex in $\boldsymbol{\alpha}$ and $\mathbf{y}$. It can also be easily verified that constraints (16) are jointly convex in $\boldsymbol{\alpha}$ and $\mathbf{y}$. Because the remaining constraints (4), (5), and (6) are linear, the reformulated DAP leads to a convex optimization problem where any local optimum solution is also a global optimum (Bazaraa and Shetty 1979). Standard convex optimization algorithms can thus be used to solve effectively for an optimal solution.

Regarding the DPP, when the first term in the objective function (10) is omitted (e.g., $h_{ij} = b_{ij} = 0 \ \forall i, j$), the DPP reduces to the GAP, which is known to be NP-hard (Fisher 1981). Therefore, the DPP is also NP-hard. To solve the problem, we use a branch-and-bound (B&B) algorithm that takes advantage of the structure of the problem and available tight lower and upper bounds to reach an optimal solution efficiently. The B&B algorithm makes use of two lower bounds. The first lower bound is given by the solution to the DAP (i.e., a relaxation of the integrality constraints). This bound is usually tight because the solution to the DAP tends to be integral (in §5 we provide a theoretical basis for this result). The second lower bound is obtained by using a dual Lagrangian relaxation of constraint set (4). The bound obtained by solving this Lagrangian relaxation is generally tighter than the one obtained via the DAP relaxation. The solution to the Lagrangian problem is also useful in guiding the branching process in the branch-and-bound algorithm. Details regarding the solution method to the dual-Lagrangian relaxation problem and the branch-and-bound algorithm can be found in Benjaafar et al. (2003). They are omitted here for the sake of brevity.

For the DAP-C, solving the joint demand allocation and inventory control problem is difficult because the problem lacks the structure that could allow us to transform it into a convex optimization problem. The problem could be solved approximately by discretizing the set of feasible values that the variables α_{ij} can assume and then solving the resulting combinatorial optimization problem. For small problems, a solution can be obtained via enumeration. For larger ones, some form of decomposition would be needed. This may be achieved, for example, by restricting the set of facilities to which a product can be assigned. In practice this might arise naturally because the set of feasible facilities for each product tends to be relatively small.

5. Numerical Results

In this section, we present numerical results that compare the solutions of the DAP, DPP, and DAP-C. We observe that while the DAP solution tends to be naturally integer, the DAP-C solution tends to split the demand from each product among multiple facilities. We refer to the tendency of the DAP solution of being integer as the *pooling effect* and the tendency of the DAP-C solution of being fractional as the *multisourcing effect* (in §6, we provide an analytical basis for both effects and provide intuition

Table 1 Comparing the DAP and DDP Solutions

		Percentage gap in cost				Percentage difference in demand allocation				Percentage of integer solutions			
(M, N)	$a = h_{ij}/c_{ij}$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.95$
(2,3)	0.05	0.2	1.395	2.187	8.534	0.397	4.608	1.763	8.072	96.7	86.7	80	66.7
	0.10	0.000	1.979	3.258	5.541	0.000	4.08	1.578	1.613	100	90	83.3	73.3
	0.15	0.000	2.581	4.132	9.414	0.000	4.062	1.584	4.902	100	90	83.3	73.3
	0.20	0.000	3.078	4.800	7.717	0.000	4.053	1.821	4.604	100	90	83.3	76.7
	0.25	0.000	3.494	5.364	9.352	0.000	4.047	1.59	1.308	100	90	83.3	76.7
	0.30	0.000	3.848	5.816	8.768	0.000	4.043	1.681	1.007	100	90	83.3	80.0
(2,5)	0.05	0.000	0.000	0.457	0.996	0.000	0.000	0.619	2.654	100	100	93.3	83.3
	0.10	0.000	0.000	0.515	2.309	0.000	0.000	0.511	2.661	100	100	96.7	83.3
	0.15	0.000	0.000	0.719	3.13	0.000	0.000	0.26	0.702	100	100	96.7	83.3
	0.20	0.000	0.000	0.883	2.356	0.000	0.000	0.266	0.549	100	100	96.7	86.7
	0.25	0.000	0.000	1.023	1.479	0.000	0.000	0.27	0.548	100	100	96.7	86.7
	0.30	0.000	0.000	1.133	1.806	0.000	0.000	0.568	0.486	100	100	96.7	86.7
(2,7)	0.05	0.000	0.000	0.000	0.448	0.000	0.000	0.000	1.454	100	100	100	93.3
	0.10	0.000	0.000	0.000	0.545	0.000	0.000	0.000	1.334	100	100	100	96.7
	0.15	0.000	0.000	0.000	0.841	0.000	0.000	0.000	1.335	100	100	100	96.7
	0.20	0.000	0.000	0.000	0.753	0.000	0.000	0.000	0.189	100	100	100	96.7
	0.25	0.000	0.000	0.000	0.844	0.000	0.000	0.000	0.189	100	100	100	96.7
	0.30	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	100	100	100	100
(4,5)	0.05	0.241	0.557	3.188	5.222	0.792	1.422	8.077	4.555	96.7	83.3	60.0	33.3
	0.10	0.000	1.105	1.734	0.038	0.000	0.363	4.086	2.912	100	93.3	56.7	33.3
	0.15	0.000	0.652	6.738	4.975	0.000	0.769	4.802	2.744	100	90.0	60.0	33.3
	0.20	0.000	0.931	8.534	7.298	0.000	0.355	3.494	2.672	100	93.3	60.0	33.3
	0.25	0.000	1.261	0.474	9.906	0.000	0.368	2.984	3.292	100	93.3	60.0	33.3
	0.30	0.000	0.189	7.257	1.673	0.000	0.143	3.691	2.507	100	96.7	60.0	36.7
(4,7)	0.05	0.000	0.000	5.489	5.422	0.000	0.000	1.300	2.564	100	100	80.0	26.7
	0.10	0.000	0.000	3.171	7.398	0.000	0.000	2.974	2.244	100	100	76.7	26.7
	0.15	0.000	0.000	1.014	9.059	0.000	0.000	0.976	2.097	100	100	83.3	33.3
	0.20	0.000	0.000	2.948	2.305	0.000	0.000	1.167	2.074	100	100	83.3	33.3
	0.25	0.000	0.000	2.158	0.063	0.000	0.000	1.047	1.899	100	100	86.7	36.7
	0.30	0.000	0.000	3.218	0.920	0.000	0.000	1.069	1.944	100	100	86.7	36.7
(4,9)	0.05	0.000	0.000	0.213	1.749	0.000	0.000	0.394	1.812	100	100	93.3	33.3
	0.10	0.000	0.000	0.274	2.517	0.000	0.000	0.258	1.526	100	100	96.7	50.0
	0.15	0.000	0.000	0.000	3.296	0.000	0.000	0.000	1.447	100	100	100	50.0
	0.20	0.000	0.000	0.262	3.179	0.000	0.000	0.182	1.139	100	100	96.7	63.3
	0.25	0.000	0.000	0.315	3.750	0.000	0.000	0.182	1.129	100	100	96.7	63.3
	0.30	0.000	0.000	0.363	4.135	0.000	0.000	0.182	1.135	100	100	96.7	63.3

for the differences between the DAP and DAP-C solutions).

In Table 1, we present representative numerical results comparing the DAP and DPP solutions. For each combination of parameters considered, we compute the percentage gap in cost, percentage difference in demand allocation, and percentage of integer solutions. The percentage gap in cost represents the relative gap of the DPP's optimal cost to that of the DAP. The percentage difference in demand allocation compares both solutions and measures how far the DAP's solution is from the DPP's solution. The percentage of integer solutions measures the percentage of times the solution of the DAP is feasible to the DPP, and hence is optimal to the DPP. Letting $\alpha^A = \{\alpha_{ij}^A\}$ and z^A ($\alpha^P = \{\alpha_{ij}^P\}$ and z^P) denote the optimal solution and the optimal cost of the DAP (the DPP), we have

$$\text{percentage gap in cost} = [(z^P - z^A)/z^P] \times 100, \text{ and}$$

$$\text{percentage difference in demand allocation} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{M} |\alpha_{ij}^A - \alpha_{ij}^P| \Big/ 2N \right\} \times 100.$$

The parameter $\rho = \sum_{i=1}^{N} \lambda_i / \sum_{j=1}^{M} \mu_j$ is used to denote the overall level of utilization in the system and is varied from 0.7 to 0.95. The ratio h_{ij}/c_{ij} is used to indicate the relative difference in inventory holding to production cost. The service level $\theta_{ij} = b_{ij}/(b_{ij} + h_{ij})$ is held fixed (in the examples shown $\theta_{ij} = 0.98$ for all values of i and j). Hence, for each value of h_{ij} there is a corresponding backorder cost b_{ij}. For ease of comparison, we also let $h_{ij}/c_{ij} = a$ for all i and j and vary the

parameter a. Each entry in Table 1 is an average over 30 randomly generated problems. Additional details on how the data are generated are in Appendix 3.

Table 1 reveals two interesting results. First, the pooling effect[1] increases as the overall utilization of the system decreases. Second, the pooling effect increases with the number of products when the number of facilities is maintained constant. The first result can be explained as follows. For lightly utilized systems, there is enough capacity to pool demand in a single location. As system loading increases, the demand rates become relatively larger and hence have to be split among facilities to reduce congestion across the system. The second result can be explained as follows. For fixed overall system utilization, when the number of products is larger than the number of facilities, the demand rate per product tends to be smaller compared to the facilities' capacities, hence the system has enough capacity to pool inventories in single locations. Note that because the solution to the DAP tends to be either integer or *nearly* integer, the DAP solution is either an optimal solution to the DPP or provides a strong lower bound to the DPP.

Note that the pooling effect is also sensitive to the ratio a. As a decreases the percentage of integer solutions also decreases. That is, the pooling effect is weakest when the holding cost, relative to the production cost, is low. This is not surprising because production costs in that case dominate and the need to consolidate inventory becomes less important. As expected, the effect of the ratio a is most important when ρ is relatively high.

In Table 2, we provide numerical results that compare the DAP and DAP-C solutions. Each entry represents an average of 30 randomly generated examples. The percentage gap in cost represents the percentage cost difference between the two systems and is computed as $[(z^A - z^C)/z^A] \times 100$, where z^A and z^C represent, respectively, the optimal cost of the DAP and DAP-C solutions. The percentage difference in demand allocation distance is computed as $\sum_{i=1}^{N}\sum_{j=1}^{M}|\alpha_{ij}^C - \alpha_{ij}^A|/2N \times 100$, where α_{ij}^C and α_{ij}^A denote, respectively, the optimal demand allocation for the DAP-C and DAP. The percentage of integer solutions measures the percentage of time an integral solution is obtained by the DAP-C.

The numerical results suggest that demand allocations with centralized inventory locations can be very different from those with factory-based inventory locations. In particular, the solution does not seem to favor the pooling of production. In fact, as indicated by the low percentage of integral solutions, there appears to be an incentive to distribute production among multiple facilities. Although the demand allocations for the two systems are very different, the cost difference (for the range of parameters shown) appears small. This result should, however, be interpreted carefully because experiments with systems with higher values of ρ show that this cost difference increases and can be arbitrarily large as ρ approaches one. As we discuss in §6.2, the ability of the DAP-C to split demand among different facilities while keeping inventory pooled allows it to achieve an additional form of risk pooling by spreading production among multiple suppliers. We provide supporting analytical arguments for this multisourcing effect.

Table 2 Comparing the DAP and DAP-C Solutions

(M, N)	ρ	Percentage gap in cost	Percentage difference in demand allocation	Percentage of integer solutions
(2,3)	0.80	1.728	19.35	13.33
	0.90	3.132	30.045	6.67
	0.95	3.915	32.878	6.67
(2,5)	0.80	0.271	5.124	26.67
	0.90	1.584	19.282	20.00
	0.95	1.899	24.309	13.33
(2,7)	0.80	0.179	3.04	33.33
	0.90	0.418	6.624	23.33
	0.95	1.162	18.25	16.67
(4,5)	0.80	4.016	45.821	3.33
	0.90	6.042	57.172	0.00
	0.95	7.543	61.181	0.00
(4,7)	0.80	1.595	26.228	13.33
	0.90	3.472	41.585	6.67
	0.95	4.670	48.942	3.33
(4,9)	0.80	0.848	18.661	16.67
	0.90	1.976	29.998	10.00
	0.95	2.634	39.151	6.67

6. Insights and Special Cases

Characterizing the optimal solution analytically for either the DAP or the DAP-C is difficult in general. However, broadly speaking, an optimal solution tends to balance costs due to three main sources: cost rates at each facility for each product, congestion at each facility, and number of inventory locations for each product type and at each facility. In lightly loaded systems (as $\rho \to 0$), production costs tend to dominate inventory-related costs. Our problem then reduces to minimizing $z(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\sum_{j=1}^{M} c_{ij}\alpha_{ij}\lambda_i$, subject to constraints (4), (5), and (6). This is, of course, the classical assignment problem and all the associated insights apply here. However, in general, although solutions tend to favor allocating more demand to the least costly facilities, excessive workloads at any facility

[1] Our usage of the term *pooling* here and elsewhere in the paper refers to the centralization of inventory of each product in few locations. When inventory is factory based, this also means that production facilities are dedicated to few products.

increase inventory-related costs. In fact, inventory-holding and backordering costs grow exponentially at each facility as its utilization increases. Hence, solutions tend to balance workloads among facilities, even if it requires assigning demand to facilities with higher production costs. For the DAP, this workload balancing must, however, account for benefits that arise from concentrating inventory (and consequently production) in as few facilities as possible (the inventory-pooling effect). For the DAP-C, because inventory is already pooled, workload can be distributed to balance workload among different facilities but also to spread risk among multiple supply sources (the multisourcing effect). Both inventory pooling and multisourcing can be viewed as examples of risk pooling. In what follows, we further examine and compare these two effects using analytical models of simple systems. Throughout, we limit our discussion to the DAP and DAP-C because in the DPP both inventory and production are pooled in a single location.

6.1. The Inventory-Pooling Effect

To highlight the role of inventory pooling in the DAP, we first illustrate how demand allocation can be very different in a make-to-order system, i.e., in a system where holding costs are sufficiently high so that base-stock levels are always set to zero. For simplicity, let us consider a system with a single product and M facilities. Then, our problem reduces to minimizing

$$z(\boldsymbol{\alpha}) = \sum_{j=1}^{M} \{b_j[r_j/(1-r_j)] + c_j\alpha_j\lambda\}, \tag{18}$$

subject to (4)–(6). Temporarily ignoring the bounding constraints (5) and (6) and using the Lagrange multiplier η with constraints (4), the Lagrangian problem can be written as

$$\text{minimize} \quad L(\boldsymbol{\alpha}) = z(\boldsymbol{\alpha}) + \eta\left(1 - \sum_{j=1}^{M} \alpha_j\right). \tag{19}$$

Differentiating with respect to α_j and setting the result to zero, we obtain the following condition on the optimal solution:

$$\alpha_j = \frac{1}{\lambda}\left(\mu_j - \sqrt{\frac{\lambda b_j \mu_j}{\eta - \lambda c_j}}\right). \tag{20}$$

For the special case where costs are symmetric, and assuming facilities are ordered such that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_M$, a closed form for the optimal solution can be obtained as

$$\alpha_j^* = \begin{cases} \dfrac{1}{\lambda}\left[\mu_j - \dfrac{\sqrt{\mu_j}\left(\sum_{i=1}^{M^*}\mu_i - \lambda\right)}{\sum_{i=1}^{M^*}\sqrt{\mu_i}}\right] & \text{if } j \leq M^*, \\ 0 & \text{if } j > M^*, \end{cases} \tag{21}$$

where M^* satisfies the condition

$$\sum_{i=1}^{M^*}\mu_i - \sqrt{\mu_{M^*}}\sum_{i=1}^{M^*}\sqrt{\mu_i} \leq \lambda \leq \sum_{i=1}^{M^*+1}\mu_i - \sqrt{\mu_{M^*}}\sum_{i=1}^{M^*+1}\sqrt{\mu_i}. \tag{22}$$

Results (21) and (22) are similar to those obtained by Ni and Hwang (1985) in the context of distributed computer systems and illustrate two important effects: (1) facilities with higher processing rates are assigned proportionally more demand, and (2) facilities with sufficiently slow processing rates are assigned no demand at all. This is an example of the so-called *dropout* rule where, depending on the demand rate, it may be optimal to idle certain facilities. Note, however, that this occurs only when facilities are heterogeneous in their capacity. In systems where facilities have equal production rates, it is always optimal to assign equal loads to all facilities, i.e., $\alpha_j^* = 1/M$ for $j = 1, \ldots, M$.

In contrast to the above result, we now show that a balanced allocation of demand among facilities in a make-to-stock system (under the DAP assumptions) even when these facilities are identical is not optimal. In particular, for a system with one product and M identical facilities, the optimal cost when demand allocation is balanced among the M facilities is given by

$$z_b^* = \frac{Mh\ln[h/(h+b)]}{\ln[\lambda/M\mu]} + \lambda c.$$

On the other hand, if all the demand is assigned to only $M-K$ facilities, the optimal cost is given by

$$z_u^* = \frac{(M-K)h\ln[h/(h+b)]}{\ln[\lambda/(M-K)\mu]} + \lambda c.$$

It is straightforward to show that $z_u^* \leq z_b^*$ holds if and only if $(M-K)(M-K)\ln[\lambda/M\mu] \leq M\ln[\lambda/(M-K)\mu]$, or equivalently $(\lambda/M\mu)^{M-K} \leq (\lambda/(M-K)\mu)^M$. The last inequality simplifies to $\lambda \leq M\mu(\sqrt[K]{[(M-K)/M]^M})$. When $K = M-1$, all demand is allocated to a single facility and the inequality reduces to $\lambda \leq \mu(\sqrt[M-1]{1/M})$.

Observation 1. In a make-to-order system with M identical facilities and one product, allocating the demand equally among the facilities is optimal. In contrast, in a make-to-stock system, where inventory is held at the facility that produces it, allocating demand to only a subset of the facilities, or even to only one facility, can be more desirable.

Observation 1 illustrates important differences between make-to-order and make-to-stock systems and shows the power of inventory pooling. By centralizing production in a single facility, inventory

can be sufficiently reduced to mitigate the increase in replenishment lead times due to higher loading and increased congestion. The result also shows that in a make-to-stock system adding capacity may not always be useful.

Next, we examine parameters that might affect the value of inventory pooling. We do so using an example of a system with N identical products and N identical facilities. Let z_d be the optimal cost when the demand from each product is equally allocated among the N facilities, and z_p be the optimal cost when the demand from each product is assigned to only one facility and each facility is assigned to only one product. Then,

$$z_d = N\left(\frac{Nh\ln[h/(h+b)]}{\ln[\rho/(N(1-\rho)+\rho)]} + \lambda c\right) \quad \text{and}$$

$$z_p = N\left(\frac{h\ln[h/(h+b)]}{\ln[\rho]} + \lambda c\right),$$

where $\rho = \lambda/\mu$, and h, b, and c are, respectively, the backorder, holding, and production costs at the facilities. It is not difficult to show that $z_p \le z_d$. Moreover, we can show that the difference $z_d - z_p$ is increasing in the parameters ρ, N, and b. Interestingly, the difference as a function of ρ and for fixed N, h, and b is bounded. Specifically, we have $\lim_{\rho\to 1} z_d - z_p = -hN(N-1)\ln[h/(h+b)]$. Although the difference is increasing in ρ, the ratio is not monotonic in ρ. It initially increases then decreases, with $\lim_{\rho\to 0} z_d/z_p = 1$ and $\lim_{\rho\to 1} z_d/z_p = 1$. Hence, the *relative* benefit of inventory pooling is most significant when utilization is in the midrange, but is limited when utilization is either low or high. When utilization is low, the result agrees with intuition, because total cost is dominated by production costs. When utilization is high, the intuition is less obvious. A plausible explanation is that positive correlation among the lead times of items being pooled in the single location increases as ρ increases, which in turn acts to reduce the relative benefit from pooling; see Benjaafar et al. (2004a) for a related discussion. The effect of utilization is summarized in the following observation.

Observation 2. For a system with N identical facilities and N products, allocating the demand from each product to only one facility is superior to allocating the demand from each product equally among the N facilities. The absolute difference in the optimal costs of the two allocations increases in utilization, but has a finite bound. In contrast, the relative benefit of pooling is decreasing in utilization when utilization is sufficiently high and eventually vanishes as utilization approaches one.

6.2. The Multisourcing Effect

To gain insights into the effect of centralizing inventory in a single location in the DAP-C on demand allocation, consider a system with two products and two facilities with identical arrival rates λ and service rates μ and identical backorder and holding costs, h and b, respectively. The expressions for expected inventory and backorder levels for each product i $(i = 1, 2)$ in Lemma 2 simplify to

$$E(I_i) = \begin{cases} s_i(1+r_i^{s_i+1}) - \dfrac{2r_i(1-r_i^{s_i})}{1-r_i} & \text{if } r_{i1} = r_{i2} = r_i, \quad \text{and} \\ \dfrac{(1-r_{i1})(1-r_{i2})}{r_{i1}-r_{i2}}\left(\dfrac{sr_{i2}(1-r_{i2}) - r_2^2(1-r_{i2}^s)}{(1-r_{i2})^2}\right. & \\ \qquad \left. - \dfrac{sr_{i1}(1-r_{i1}) - r_{i1}^2(1-r_{i1}^s)}{(1-r_{i1})^2}\right) & \text{otherwise,} \end{cases} \tag{23}$$

and

$$E(B_i) = \begin{cases} s_i r_i^{s_i+1} + \dfrac{2r_i^{s_i+1}}{1-r_i} & \text{if } r_{i1} = r_{i2} = r_i, \text{ and} \\ \dfrac{(1-r_{i1})(1-r_{i2})}{r_{i1}-r_{i2}}\left(\dfrac{r_{i2}^{s_i+2}}{(1-r_{i2})^2} - \dfrac{r_{i1}^{s_i+1}}{(1-r_{i1}^2)}\right) & \\ \qquad \text{otherwise.} \end{cases} \tag{24}$$

We consider the following two demand allocation scenarios. In Scenario 1, all of Product 1 demand is assigned, to Facility 1, and all of Product 2 demand is assigned to Facility 2. In Scenario 2, demand from Products 1 and 2 are split evenly among Facilities 1 and 2. That is, each facility is assigned half the demand of each product. In both scenarios, the utilization of the facilities is the same $\rho = \lambda/\mu$. We refer to Scenario 1 as the pooled scenario (both production and inventory are pooled in a single location) and Scenario 2 as the distributed scenario (inventory is pooled but production is distributed). In Observation 3, we show that the distributed scenario is more desirable than the pooled one and that the difference between the two scenarios becomes infinitely large as ρ approaches one.

Observation 3. Let z_p^* and z_d^* denote the optimal cost under the pooled and distributed scenarios, respectively. Then, (a) $z_d^* \le z_p^*$, (b) $\lim_{\rho\to 1}(z_p^* - z_d^*) = \infty$, and (c) $\lim_{\rho\to 0}(z_p^* - z_d^*) = 0$.

Proof. Expected cost (given base-stock levels s_1 and s_2 for Products 1 and 2, respectively) under scenario j (j = pooled, distributed) can be written as

$$z_j(s_1, s_2) = 2c\lambda + \sum_{i=1}^{2}(h(s_i - E(Q_{i,j})) + (h+b)E(B_{i,j})).$$

Because $E(Q_{i,d}) = \sum_{j=1}^{2} r_{i,j}/(1-r_{i,j})$ and $r_{i,j} = \rho/(2-\rho)$ for $i, j = 1, 2$, we have $E(Q_{i,d}) = \rho/(1-\rho)$. However, $E(Q_{i,p}) = \rho/(1-\rho)$ and, consequently, we have $E(Q_{i,p}) = E(Q_{i,d}) = \rho/(1-\rho)$. To show that $z_p(s_1, s_2) \ge z_d(s_1, s_2)$ for any combination of s_1 and s_2,

it then suffices to show that $E(B_{i,p}) \geq E(B_{i,d})$, or equivalently, that the ratio $\delta_B = E(B_{i,d})/E(B_{i,p}) \leq 1$. To show the latter we note that $\delta_B = [s_i(1-\rho)+2-\rho]/(2-\rho)^{s_i+1}$. Because $\delta_B = (s_i+2)/2^{s_i+1} < 1$ when $\rho = 0$, $\delta_B \to 1$ as $\rho \to 1$, and δ_B is strictly increasing in ρ, we have indeed $\delta_B \leq 1$. If we let $s^*_{i,p}$ and $s^*_{i,d}$ denote the optimal base-stock levels for product i under the pooled and distributed scenarios, respectively, then we have $z_p(s^*_{1,p}, s^*_{2,p}) \geq z_d(s^*_{1,p}, s^*_{2,p}) \geq z_d(s^*_{1,d}, s^*_{2,d})$, which proves Part (a). For Part (b), first note that the optimal stock level and the optimal cost under the pooled scenario are given by

$$s^*_{i,p} = \frac{\ln(h/(h+b))}{\ln(\rho)} \quad \text{for } i = 1,2 \quad \text{and}$$

$$z^*_p = 2\left\{c\lambda + h\frac{\ln(h/(h+b))}{\ln(\rho)}\right\}.$$

Let $s^*_p = s^*_{i,p}$ for $i = 1,2$. Then, the difference $z_p(s^*_p, s^*_p) - z_d(s^*_p, s^*_p)$ is given by

$$z_p(s^*_p, s^*_p) - z_d(s^*_p, s^*_p) = 2h\left(\frac{\rho}{1-\rho}\right) - 2(b+h)\rho r^{s^*_p}\left(\frac{s^*_p}{2-\rho} + \frac{1}{1-\rho}\right).$$

Taking the limit and applying l'Hopital's rule leads to

$$\lim_{\rho\to 1}\left(z_p(s^*_p, s^*_p) - z_d(s^*_p, s^*_p)\right) = \infty.$$

Because $z_d(s^*_p, s^*_p) \geq z_d(s^*_d, s^*_d)$, we have

$$\lim_{\rho\to 1}(z_p(s^*_p, s^*_p) - z_d(s^*_d, s^*_d)) = \infty.$$

Finally, for Part (c), we note that as $\rho \to 0$, $s^*_p = s^*_d = 0$ and $z^*_p = z^*_d = 2c\lambda$. Hence, $\lim_{\rho\to 0}(z^*_p - z^*_d) = 0$. □

Observation 3 shows that given that the inventory is pooled, it is beneficial to allocate production among multiple facilities. This does make intuitive sense. By distributing production among more than one facility, variability associated with the supply process is reduced by spreading it over several sources. This multisourcing of production can be seen as another form of hedging, or risk pooling, available to systems with multiple facilities. In Observation 4, we verify the validity of this intuition. We show that the variance of inventory on order (this is also *lead-time demand* in our setting) is reduced under the distributed scenario and that the difference in variance grows infinitely large with utilization.

Observation 4. Let $\mathrm{Var}(Q_{i,p})$ and $\mathrm{Var}(Q_{i,d})$ denote the variance of the number of units of type i on order under the pooled and distributed scenarios, respectively. Then, $\Delta_v = \mathrm{Var}(Q_{i,p}) - \mathrm{Var}(Q_{i,d}) = \rho^2/2(1-\rho)^2$ and $\delta_v = \mathrm{Var}(Q_{i,p})/\mathrm{Var}(Q_{i,d}) = 2/(2-\rho)$. Then, Δ_v is increasing in ρ, with $\Delta_v \geq 0$, $\lim_{\rho\to 0}\Delta_v = 0$, $\lim_{\rho\to 1}\Delta_v = \infty$, $\lim_{\rho\to 0}\delta_v = 1$, and $\lim_{\rho\to 1}\delta_v = 2$.

Proof. The results follow from noting that

$$\mathrm{Var}(Q_{i,p}) = \frac{\rho}{(1-\rho)^2} \quad \text{and}$$

$$\mathrm{Var}(Q_{i,d}) = \frac{r_{i1}}{(1-r_{i1})^2} + \frac{r_{i2}}{(1-r_{i2})^2} = \frac{2r}{(1-r)^2},$$

where the latter follows from the independence of facilities 1 and 2. Using the fact that $r = \rho/(2-\rho)$, we obtain $\mathrm{Var}(Q_{i,d}) = \rho(2-\rho)/2(1-\rho)^2$, which after some algebra leads to the desired results. □

The effect of the variance on the optimal cost can be made more visible when we approximate the distribution of inventory on order by a normal distribution with matching moments. This approximation also allows us to consider for the above example the more general case of N products and N facilities. Under the normal approximation, we approximate the distributions of $Q_{i,p}$ and $Q_{i,d}$ by normal distributions with means $E(Q_{i,p}) = \rho/(1-\rho)$ and $E(Q_{i,d}) = Nr/(1-r) = \rho/(1-\rho)$ and variances $\mathrm{Var}(Q_{i,p}) = \rho/(1-\rho)^2$ and $\mathrm{Var}(Q_{i,d}) = Nr/(1-r)^2$. The optimal costs for a system with N products and N facilities can then be in turn approximated as

$$z^*_p \approx \hat{z}^*_p = cN\lambda + Nk\sqrt{\mathrm{Var}(Q_{i,p})} = c\lambda + Nk\sqrt{\rho/(1-\rho)^2}$$

and

$$z^*_d \approx \hat{z}^*_d = cN\lambda + Nk\sqrt{\mathrm{Var}(Q_{i,d})} = c\lambda + Nk\sqrt{Nr/(1-r)^2},$$

where k is a constant that depends on the ratio $b/(h+b)$. As discussed in Chapter 7 of Zipkin (2000), the normal approximation can be shown to retain the qualitative characteristics of the optimal cost under the exact distribution. The difference $\hat{z}^*_p - \hat{z}^*_d$ is given by

$$\hat{z}^*_p - \hat{z}^*_d = \sqrt{\frac{\rho}{(1-\rho)^2}\left(\frac{N}{N(1-\rho)+\rho}\right)},$$

which is clearly increasing in ρ, with $\lim_{\rho\to 0}\hat{z}^*_p - \hat{z}^*_d = 0$ and $\lim_{\rho\to 1}\hat{z}^*_p - \hat{z}^*_d = \infty$. The ratio $\hat{z}^*_p/\hat{z}^*_d$ is also increasing in ρ, with $\lim_{\rho\to 0}\hat{z}^*_p/\hat{z}^*_d = 1$ and $\lim_{\rho\to 1}\hat{z}^*_p/\hat{z}^*_d = \sqrt{N}$.

The above analysis highlights how optimal costs are sensitive to the variance of inventory on order. By splitting demand, variance can be significantly reduced and hence, the optimal cost similarly reduced. This reduction is particularly significant when utilization is high. In the limit case the optimal cost is reduced by a factor of approximately $\sqrt{N}$.

The risk pooling achieved with multisourcing is similar to the one achieved from inventory pooling when inventory is factory based, although there are some important differences. In contrast to multisourcing, the relative benefit from inventory pooling

decreases in ρ and eventually vanishes as $\rho \to 1$. Also, the absolute benefit from inventory pooling remains bounded as ρ increases while the absolute benefit from multisourcing grows without bound. This difference in the effect of ρ on the two forms of risk pooling might be related to the fact that when inventory is pooled as in the DAP the demand streams that are pooled are served from a single production facility. In that case, higher utilization induces higher correlation between the lead time demands of these streams. In contrast, with centralized inventory, higher utilization does not affect correlation because the demand streams being split are served from independent facilities.

The inventory pooling and multisourcing effects highlight the need for firms to coordinate their distribution strategy with their choice of manufacturing technology. The value of flexible manufacturing technology is greater when inventory is centrally warehoused. On the other hand, there is less justification for investments in flexible technology if inventories are warehoused at the factories. The results also illustrate the value of a dual strategy of "pooling demand" via the centralization of inventory but "splitting supply" via multisourcing.

7. Extensions

In this section, we present extensions to our original model. In particular, we consider systems with multiple customer classes, systems with heterogeneous service times, and systems with general distributions of demand and service times.

7.1. Systems with Multiple Customer Classes

In certain settings, the demand for a particular product may arise from different customer classes. Different backordering costs (or service levels) may be associated with different customer types. For the DAP, in addition to determining which products should be assigned to which facility, there is now a need to determine from which facility the demand for each customer class should be satisfied. In other words, our demand decision variables for the DAP are the parameters α_{ij}^k, where α_{ij}^k refers to the fraction of demand of customer k for product i that is satisfied from facility j. The α_{ij}^ks must satisfy the constraint

$$\sum_{j=1}^{M} \alpha_{ij}^k = 1, \quad i = 1, 2, \ldots, N, \ k = 1, 2, \ldots, K_i, \qquad (25)$$

and K_i is the total number of customer classes for product i.

Due to the differences in backorder costs of customers, backordering an arriving demand for product i at facility j, in consideration of future more expensive arrivals, can reduce the long-run total cost. Recently, de Véricourt et al. (2002) established a complete characterization of the rationing policy minimizing the long-run average inventory costs for a single product at a single location. In that case, there is a threshold for each class of customers such that it is optimal to satisfy demands from this class if the on-hand inventory is above that threshold. More precisely, consider n thresholds $s_1, \ldots, s_n$, where s_n corresponds to the base-stock level of the system and s_{k-1} is the threshold associated with customer k. Then, the optimal rationing policy is of the form:

- allocate a produced item to a waiting demand of customer class k if the stock level is at s_{k-1}, and
- backorder an arriving demand from customer class k if the inventory level is less than s_{k-1}.

The optimal thresholds and the optimal base-stock level $(s_1^*, \ldots, s_n^*)$, as well as the corresponding optimal long-run average inventory cost, can be computed recursively.

When more than one product is considered, because the allocation of the production capacity among the different product types follows a (FCFS) policy, the distribution of inventory on order at facility j, $Q_j = \sum_{i=1}^{N} Q_{ij}$, where Q_{ij} is the number of orders of product i placed with facility j, has a product form distribution with Q_{ij} being geometrically distributed with parameter r_{ij} (see Buzacott and Shanthikumar 1993 for details). The distribution of Q_{ij} is also the one we obtain in a facility with a single product but service rate $\mu_{ij} = \mu_j - \sum_{l \neq i} \sum_{k=1}^{K_l} \alpha_{lj}^k \lambda_l^k$. The rationing problem for N products at facility j with production rate μ_j is then equivalent to N rationing subproblems with production rates μ_{ij}. Hence, we can apply the recursive procedure presented in de Véricourt et al. (2002, Property 3) to compute the total expected cost $z(\boldsymbol{\alpha}, \mathbf{s}^*)$, where $\mathbf{s}^*$ is a matrix with components s_{ij}^k, representing the inventory threshold for customer k of product i at facility j. Because the state space is discrete, this procedure provides integer thresholds. Treating these thresholds as real values, we can obtain the following much simpler expression for the optimal cost:

$$z(\boldsymbol{\alpha}, \mathbf{s}^*) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{K_i} \left\{ h_{ij} \ln \left[\frac{h_{ij} + b_{ij}^{k+1}}{h_{ij} + b_{ij}^k} \right] \cdot \frac{1}{\ln[\sum_{l=1}^{k} \hat{r}_{ij}^l]} + c_{ij} \alpha_{ij}^k \lambda_i^k \right\}, \qquad (26)$$

where b_{ij}^k is the backordering cost for customer class k for product i at facility j, with $b_{ij}^{K_i+1} = 0$, λ_i^k is the demand rate of customer k for product i, and

$$\hat{r}_{ij}^k = \frac{\alpha_{ij}^k \lambda_i^k}{\mu_j - \sum_{l \neq i} \sum_{t=1}^{K_l} \alpha_{lj}^t \lambda_l^t}. \qquad (27)$$

The DAP can now be solved by first solving the following related problem with decision variables x_{ij}^k and y_{ij}^k:

minimize

$$\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{c_{ij}x_{ij}^{K_i}-\sum_{k=1}^{K_i}\frac{h_{ij}\ln[(h_{ij}+b_{ij}^{k+1})/(h_{ij}+b_{ij}^{k})]}{y_{ij}^{k}}\right\} \quad (28)$$

subject to

$$\sum_{j=1}^{M}x_{ij}^{k}=\sum_{t=1}^{k}\lambda_i^t, \quad i=1,2,\ldots,N,\ k=1,2,\ldots,K_i, \quad (29)$$

$$\sum_{i=1}^{N}x_{ij}^{K_i}-\mu_j\le 0, \quad j=1,2,\ldots,M, \quad (30)$$

$$x_{ij}^{k}-x_{ij}^{k-1}\ge 0, \quad i=1,2,\ldots,N,\ j=1,2,\ldots,M,\ k=1,2,\ldots,K_i, \quad (31)$$

$$x_{ij}^{k}e^{y_{ij}^{k}}+\sum_{l\neq i}x_{lj}^{K_l}-\mu_j\le 0, \quad i=1,2,\ldots,N,\ j=1,2,\ldots,M,\ k=1,2,\ldots,K_i, \quad (32)$$

$$x_{ij}^{1}\ge 0, \quad y_{ij}^{k}\ge 0, \quad i=1,2,\ldots,N,\ j=1,2,\ldots,M,\ k=2,\ldots,K_i. \quad (33)$$

Both the objective function and the constraints in the above problem are jointly convex in the decision variables. Once we solve the above problem, the original decision variables can be recovered as follows: $\alpha_{ij}^k = (x_{ij}^k - x_{ij}^{k-1})/\lambda_{ij}^k$ with $\alpha_{ij}^1 = x_{ij}^1/\lambda_{ij}^1$. It is easy to verify that the α_{ij}^ks satisfy all the constraints in the original problem.

Although inventory rationing in systems with multiple customer classes is a superior policy to an inventory allocation based on FCFS, it can be difficult or costly to implement in practice. Therefore, it is of interest to examine the extent to which optimal rationing improves performance over FCFS and how different the demand allocations are under the two policies. Under FCFS, the total cost is given by

$$z(\boldsymbol{\alpha},\mathbf{s}^*)=\sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\frac{h_{ij}\ln[h_{ij}/(h_{ij}+\hat{b}_{ij})]}{\ln[\hat{r}_{ij}]}+c_{ij}\sum_{k=1}^{K_i}\alpha_{ij}^k\lambda_i^k\right\}, \quad (34)$$

where

$$\hat{r}_{ij}=\frac{\sum_{k=1}^{K_i}\alpha_{ij}^k\lambda_i^k}{\mu_j-\sum_{l\neq i}\sum_{k=1}^{K_i}\alpha_{lj}^k\lambda_l^k} \quad \text{and} \quad \hat{b}_{ij}=\frac{\sum_{k=1}^{K_i}\alpha_{ij}^k\lambda_{ik}b_{ij}^k}{\sum_{k=1}^{K_i}\alpha_{ij}^k\lambda_i^k}. \quad (35)$$

The optimal base-stock level for each product at each facility (for a given flow allocation matrix $\boldsymbol{\alpha}$) can be obtained as

$$s_{ij}^*=\ln[h_{ij}/(h_{ij}+\hat{b}_{ij})]/\ln[\hat{r}_{ij}]. \quad (36)$$

Unfortunately, the problem under FCFS cannot be easily transformed into a convex optimization problem because of the definition of $\hat{b}_{ij}$ in (35). However, in all of our numerical examples, a local search algorithm has been converging to a unique optimal solution.

We have so far considered systems where inventory is held locally at each facility. The analysis can be extended to systems where inventory for each product is centrally held. In this case, a rationing threshold policy means that demand from customer k for product i is backordered if inventory level for product i in the central location is less than s_i^k; otherwise it is satisfied from available stock. Furthermore, a produced item is allocated to a waiting demand of customer class k if the stock level is at the corresponding threshold s_i^{k+1}. A demand from class k for product i always triggers a production order that is allocated to facility j with probability α_{ij}. Under the above assumptions, it can be shown that the optimal cost is given by (see Appendix 4 for supporting arguments)

$$z(\boldsymbol{\alpha},\mathbf{s}^*)=\sum_{i=1}^{N}\left[h_i s_i^*+\sum_{j=1}^{M}c_{ij}\alpha_{ij}\lambda_{ij}\right], \quad (37)$$

where the optimal base-stock levels s_i^* are recursively computed as follows: (1) let $s_i^0 = 0$, (2) for $0 \le k \le K_i$, let s_i^{k+1} be the unique solution to

$$\sum_{j=1}^{M}R_{ij}^k(\hat{r}_{ij}^k)^{s-s_i^k}=\frac{h_i+b_i^{k+1}}{h_i+b_i^k}, \quad \text{where}$$

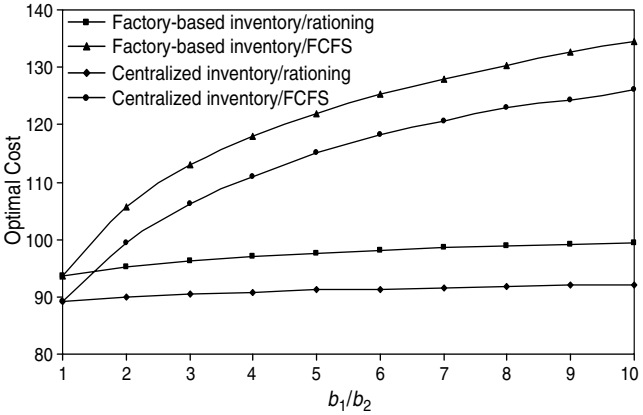
$$R_{ij}^k=(\hat{r}_{ij}^k)^{M-1}\prod_{l\neq j}\frac{(1-\hat{r}_{il}^k)}{(\hat{r}_{ij}^k-\hat{r}_{il}^k)}, \quad (38)$$

and (3) let $s_i^* = s_i^{K_i+1}$, where s_i^{k+1} corresponds to the optimal threshold rationing levels for customer k for product i.

If instead of a rationing policy we use a FCFS policy to allocate demand, then the problem is equivalent to the DAP-C problem with λ_i and b_i substituted by $\sum_{k=1}^{K_i}\lambda_i^k$ and $\hat{b}_i = \sum_{k=1}^{K_i}\lambda_i^k b_i^k/\lambda_i$, respectively.

To examine the magnitude of cost savings from inventory rationing, let us consider a system with two identical facilities, one product, and two customer classes. The two customer classes have the same demand rates but different backorder costs b_1 and b_2. In Figure 1, we illustrate the effect of increasing the ratio b_1/b_2 on the optimal cost for four system configurations: (a) centralized inventory with optimal rationing, (b) centralized inventory with FCFS allocation, (c) factory-based inventory with optimal rationing, and (d) factory-based inventory with FCFS allocation. The graphs reveal three interesting insights: (1) there can be significant value to inventory rationing regardless of whether inventory is centralized or

Figure 1 The Effect of Backorder Costs on the Optimal Cost for Different System Configurations



Note. $\rho = 0.95$, $\mu = 1$, $c = 0$, $h = 1$, and $b_2 = 10$.

not, (2) the cost savings from inventory rationing can be more significant than those from inventory centralization, and (3) the cost difference between systems without and with rationing increases with the ratio b_1/b_2. Numerical experiments with systems with larger numbers of products, facilities, and customer classes confirm these insights. These results suggest that for firms facing multiple customer classes adopting a strategy of differentiated customer service (via inventory rationing) can be more value-adding than a strategy of inventory pooling.

The demand allocations that arise under the four configurations can also be quite different (for the sake of brevity, detailed numerical results are omitted). When inventory is factory based, there are significant differences in how demand is allocated depending on whether or not inventory is rationed. If FCFS allocation is used, the demands from the two classes are largely segregated with Class 1 served mostly from one facility and Class 2 served entirely from the other facility. On the other hand, if inventory is rationed, the demand from each class is evenly split between the two facilities. Although initially surprising, the differences between the two policies do make intuitive sense. Under FCFS, segregating the two classes allows us to offer differentiated service levels to each class (i.e., choose a higher base-stock level for Class 1 than for Class 2). We call this the *customer segmentation effect*. Under rationing, differentiated service levels can be offered without physically separating the classes. Consequently, there is an incentive to mix classes in both facilities, because this gives priority to Class 1 in more than one facility. We call this the *inventory rationing effect*. We should note that the customer segmentation under FCFS is generally not complete. We have observed that there is generally a small fraction of demand from Class 1 that gets assigned to Facility 2, the facility from which all of Class 1 demand is served. This appears due to the fact that subjecting a small fraction of Class 1 demand to the lesser service quality in Facility 2 can improve the quality of service of the larger fraction that remains in Facility 1. This is, however, possible only within a limited range (a few percentage points of Class 1 demand in all of the examples we generated). When inventory is centralized, the demand allocated to production facilities is undifferentiated because regardless of the allocation all demand is met from a single inventory location. As expected, under both FCFS and rationing, demand is split evenly between the two facilities.

These results suggest that, interestingly, a policy of undifferentiated customer service (i.e., FCFS allocation) gives rise to customer-focused facilities while a policy of differentiated customer service gives rise to multicustomer facilities. In practice, firms must be aware of these effects when adopting a strategy of differentiated service. In particular, they must put in place processes that can support servicing multiple customer classes from each facility. These concerns, of course, arise only when inventory is factory based and are absent when inventory is centralized in a single location.

7.2. Products with Heterogeneous Processing Requirements and General Demand Distributions

In environments where the processing requirements of different products vary significantly, there may be a need to explicitly account for these differences in characterizing the distribution of production times at each facility. In particular, let the processing times of type i products at facility j be i.i.d. random variables, denoted by S_{ij} with probability distribution F_{ij}. Then, $Q_j = \sum_{i=1}^{N} Q_{ij}$, where Q_{ij} is the number of orders of type i placed with facility j, is equal in distribution to the number of customers in an $M/G/1$ queue with arrival rate $\Lambda_j = \sum_{i=1}^{N} \alpha_{ij}\lambda_i$ and service time probability distribution $F_j = \sum_{i=1}^{N} (\alpha_{ij}\lambda_i / \sum_{i=1}^{N} \alpha_{ij}\lambda_i) F_{ij}$, a result that holds as long as orders are processed on an FCFS basis. Then, given the stationary distribution of the number of customers in the corresponding $M/G/1$ queue, we can derive the stationary probability distribution of the Q_{ij}s in a manner analogous to the one described in Appendix 1. Unfortunately, a closed-form expression for the queue size distribution in an $M/G/1$ queue is difficult to obtain. In what follows, we propose approximating the queue size in an $M/G/1$ queue by a geometric distribution of the form

$$\Pr(Q_j = n_j) = \begin{cases} 1 - \rho_j & \text{if } Q_j = 0, \\ \rho_j(1 - \sigma_j)\sigma_j^{n_j - 1} & \text{if } Q_j = 1, 2, \ldots, \end{cases} \tag{39}$$

where

$$\sigma_j = (E(Q_j) - \rho_j)/E(Q_j),$$
$$E(Q_j) = \frac{\sum_{i=1}^{N} \alpha_{ij}\lambda_i \sum_{i=1}^{N} \alpha_{ij}\lambda_i E(S_{ij}^2)}{2(1-\rho_j)} + \rho_j, \qquad (40)$$

and $\rho_j = \sum_i \alpha_{ij}\lambda_i E(S_{ij})$. Supporting arguments for using such an approximation can be found in Buzacott and Shanthikumar (1993) and Tijms (1995). From (39)–(41), and using arguments similar to those used in Appendix 1, we obtain

$$\Pr(Q_{ij} = n_{ij}) = \begin{cases} 1 - \hat{r}_{ij} & \text{if } Q_{ij} = 0, \\ \dfrac{\rho_j}{\sigma_j}(1-\hat{r}_{ij})\hat{r}_i^{n_{ij}} & \text{if } Q_{ij} = 1, 2, \ldots, \end{cases} \qquad (41)$$

where $\hat{r}_{ij} = \alpha_{ij}\lambda_i\sigma_j/(\sum_i \alpha_{ij}\lambda_j - \sigma_j \sum_{k\neq i} \alpha_{kj}\lambda_k)$. From (41), we can now express for the DAP (a similar analysis can be carried out for the DAP-C) the expected inventory and backorder level for each item at each facility, respectively, as

$$E(I_{ij}) = s_{ij} - \frac{\rho_j}{\sigma_j}\left(\frac{\hat{r}_{ij}}{1-\hat{r}_{ij}}\right)(1-\hat{r}_{ij}^{s_{ij}}) \quad \text{and}$$
$$E(B_{ij}) = \frac{\rho_j}{\sigma_j}\left(\frac{\hat{r}_{ij}^{s_i+1}}{1-\hat{r}_{ij}}\right). \qquad (42)$$

For a given flow allocation matrix, $\boldsymbol{\alpha}$, we can then show that the optimal base-stock level is given by

$$s_{ij}^* = \ln\left[\frac{h_{ij}\sigma_j}{\rho_j(h_{ij}+b_{ij})}\right] \Big/ \ln[\hat{r}_{ij}]. \qquad (43)$$

Substituting into the objective function, we obtain

$$z(\boldsymbol{\alpha}, \mathbf{s}^*) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\frac{h_{ij}\ln[h_{ij}\sigma_j/\rho_j(h_{ij}+b_{ij})]}{\ln[\hat{r}_{ij}]} + c_{ij}\alpha_{ij}\lambda_{ij}\right\}. \qquad (44)$$

Unfortunately, this objective function does not share similar properties as (10) and (26) and is neither convex nor concave in the α_{ij}s. Hence, global optimization techniques have to be used to determine an optimal allocation. Nevertheless, insights can still be gained by examining the objective function.

From (44), we can see that the cost of the optimal demand allocation is sensitive to the value of the parameter σ_j. In particular, for fixed ρ_j, the optimal base-stock level is increasing in σ_j. The parameter σ_j, which lies on the interval [0, 1], is itself a function of two parameters: ρ_j, the utilization of facility j, and C_{s_j}, the coefficient of variation in processing times at facility j. This can be viewed more conveniently by rewriting (40) as follows:

$$E(Q_j) = \frac{\rho_j^2(1+C_{s_j}^2)}{2(1-\rho_j)} + \rho_j, \quad \text{where}$$
$$C_{s_j}^2 = \frac{\text{Var}(S_j)}{E(S_j)^2} = \frac{\sum_{i=1}^{N}\alpha_{ij}E(S_{ij}^2) - (\sum_{i=1}^{N}\alpha_{ij}E(S_{ij}))^2}{(\sum_{i=1}^{N}\alpha_{ij}E(S_{ij}))^2}. \qquad (45)$$

From (45), we can see that for fixed ρ_j, σ_j is increasing in C_{s_j}. Hence, all else being equal, the optimal solution will seek demand allocations that minimize the variability in processing times. This points to yet another principle, which we call the *processing variability effect*, that must be balanced against the other effects in determining an optimal solution.

The approximation in (41) could in principle be used to handle systems where the product demands have a general distribution, but still form independent renewal processes. However, two technical difficulties must be addressed. The first arises from the fact that the superposition of renewal processes does not necessary yield a renewal process. Therefore, the arrival process to each facility may not be renewal. The second is due to the lack of an exact expression for the expected number of customers 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, such as the asymptotic method described in Albin (1984) and Whitt (1982). The second difficulty can be addressed by using one of the many reasonably good bounds available for the expected number in system in a $GI/G/1$ queue (e.g., Wolff 1989). Alternatively, we may focus on regimes where explicit results are available. One such regime is heavy traffic. In particular, it is known that as $\rho_j \to 1$, the number of class i customers in a multiclass $G/G/1$ queue j weakly converges to a Reflected Brownian Motion with drift $\alpha_{ij}\lambda_i\rho_j^{-1}(1-\rho_j)$ and variance (Peterson 1991)

$$(\alpha_{ij}\lambda_i)^2\rho_j^{-2}\sum_{i=1}^{N}\alpha_{ij}\lambda_i E(S_{ij})^2(C_{a_{ij}}^2 + C_{s_{ij}}^2), \qquad (46)$$

where $C_{a_{ij}}$ and $C_{s_{ij}}$ are, respectively, the coefficients of variation in interarrival times and processing times for product i at facility j, with

$$C_{a_{ij}}^2 = \alpha_{ij}C_{a_i}^2 + 1 - \alpha_{ij}, \qquad (47)$$

where $C_{a_i}^2$ is the coefficient of variation in interarrival times for product i. For a given flow allocation matrix,

$\boldsymbol{\alpha}$, it can then be shown that the optimal base-stock level is given by

$$s_{ij}^* = \frac{\alpha_{ij}\lambda_i \ln[(h_{ij}+b_{ij})/h_{ij}] \sum_{i=1}^{N} \alpha_{ij}\lambda_i E(S_{ij})^2 (C_{a_{ij}}^2 + C_{s_{ij}}^2)}{2\rho_j(1-\rho_j)} \tag{48}$$

(see Wein 1992 for details). Substituting into the objective function, we obtain

$$z(\boldsymbol{\alpha}, \mathbf{s}^*) = \sum_{i=1}^{N}\sum_{j=1}^{M} \left\{ \frac{\alpha_{ij}\lambda_i h_{ij} \ln[(h_{ij}+b_{ij})/h_{ij}] \sum_{i=1}^{N} \alpha_{ij}\lambda_i E(S_{ij})^2 (C_{a_{ij}}^2 + C_{s_{ij}}^2)}{2\rho_j(1-\rho_j)} + c_{ij}\alpha_{ij}\lambda_i \right\}. \tag{49}$$

As it can be seen from (49), the optimal allocation is affected by the parameters $C_{a_{ij}}$. All else being equal, the optimal allocation will attempt to minimize variability in order arrivals to the facilities. It is interesting to note here that, depending on the amount of variability in demand processes, there may be a *disincentive* toward inventory pooling. In particular, splitting demand among multiple facilities can reduce the variability in the order arrival process to facilities. This can be seen by noting that the inequality

$$C_{a_{ij}}^2 = \alpha_{ij} C_{a_i}^2 + 1 - \alpha_{ij} \le C_{a_i}^2$$

holds if and only if $C_{a_i} \ge 1$. This means that demand splitting would reduce the arrival variability of orders to the production facilities when the variability of the original demand processes is relatively high ($C_{a_i} > 1$). This is rather counterintuitive because the benefits of inventory pooling are generally thought to be greater when demand variability is high. This also highlights an eighth principle, a *demand-variability effect*, of which managers must be aware in making demand allocation decisions.

Although we have restricted our discussion to the DAP, the analysis can be extended to the DAP-C. As in the DAP, demand splitting has the effect of reducing interarrival time variability at the facilities when the original demand variability is sufficiently high. This reduction in arrival variability compounds the benefit derived from splitting due to the multisourcing effect (i.e., the reduction in the variance of inventory on order). The two effects are different. For example, splitting demand when the coefficients of variation of demand interarrival times are less than one actually increases the variability of the arrival process, while it maintains some desirable variance-reduction effect. However, the net effect could be that demand splitting becomes undesirable in some cases (e.g., splitting demand probabilistically in a system with little or no interarrival and processing time variability).

8. Concluding Comments

In this paper, we have presented models for assisting managers in making product demand allocation and inventory control decisions in environments where there are multiple facilities capable of manufacturing each product. We used the models to illustrate how they can be used to gain insights into factors that affect the desirability of demand allocation and inventory control policies. In particular, we highlighted eight important principles that relate the effect of production cost, congestion, inventory pooling, multiple sourcing, customer segmentation, inventory rationing, and process/demand variability.

The models we presented could be used to examine additional strategic decisions. For example, the models could be embedded in a capacity-planning model to jointly determine the size and number of facilities as well as the demand allocation to these facilities. The models could also be useful in determining the optimal level of production flexibility in settings where increasing the scope of each facility requires an investment cost. Alternatively, we may use the models to examine product redesign strategies that would reduce product variety via standardization or delayed product differentiation.

Acknowledgments

Research of the first author was supported in part by the National Science Foundation under grant DMI 9988721 and by a grant from the Honeywell Corporation. The authors are grateful for useful comments from the associate editor and two anonymous reviewers.

Appendix 1. Proof of Lemma 1

The conditional probability $\Pr(Q_{ij} = n_{ij} \mid Q_j = n_j)$, where Q_{ij} is the number of units of type i that are on order with facility j and $Q_j = Q_{1j} + Q_{2j} + \cdots + Q_{Nj}$, has a binomial distribution with probability $p_{ij} = \alpha_{ij}\lambda_i / (\sum_{k=1}^{N} \alpha_{kj}\lambda_k)$. Hence, for all $n_j \ge n_{ij}$, we have

$$\Pr(Q_{ij} = n_{ij} \mid Q_j = n_j) = \frac{n_j!}{n_{ij}!(n_j - n_{ij})!} (p_{ij})^{n_{ij}} (1-p_{ij})^{n_j - n_{ij}}. \tag{A1}$$

Noting that the total numbers of orders in a facility j is equal in distribution to the number of customers in an $M/M/1$ queue, we have $\Pr(Q_j = n_j) = \rho_j^{n_j}(1-\rho_j)$, and we obtain

$$\begin{aligned}\Pr(Q_{ij} = n_{ij}) &= \sum_{n_j = n_{ij}}^{\infty} \Pr(Q_{ij} = n_{ij} \mid Q_j = n_j) \Pr(Q_j = n_j) \\ &= \sum_{n_j = n_{ij}}^{\infty} \frac{n_j!}{n_{ij}!(n_j - n_{ij})!} p_{ij}^{n_{ij}} (1-p_{ij})^{n_j - n_{ij}} \rho_j^{n_j} (1-\rho_j) \\ &\qquad \text{for } n_{ij} \ge 1, \end{aligned} \tag{A2}$$

which after some algebra leads to $\Pr(Q_{ij} = n_{ij}) = (1 - r_{ij}) r_{ij}^{n_{ij}}$, where

$$r_{ij} = \frac{\rho_j p_{ij}}{1 - \rho_j(1-p_{ij})} = \frac{\alpha_{ij}\lambda_i}{\mu_j - \sum_{k \ne i} \alpha_{kj}\lambda_k}. \tag{A3}$$

For $Q_{ij} = 0$, we have $\Pr(Q_{ij} = 0) = 1 - \sum_{n_{ij}=1}^{\infty} \Pr(Q_{ij} = n_{ij}) = 1 - r_{ij}$. We can now obtain the expected inventory for each item as $E(I_{ij}) = \sum_{n_{ij}}^{s_{ij}} (s_{ij} - n_{ij}) \Pr(Q_{ij} = n_{ij}) = s_{ij} - r_{ij}(1 - r_{ij}^{s_{ij}})/(1 - r_{ij})$, and the expected number of backorders as $E(B_{ij}) = \sum_{n_{ij}=s_{ij}}^{\infty} (n_{ij} - s_{ij}) \Pr(Q_{ij} = n_{ij}) = r_{ij}^{s_{ij}+1}/(1 - r_{ij})$. Finally, it is straightforward to show that $\Pr(Q_{ij} \geq s_{ij}) = 1 - \sum_{n_{ij}=0}^{s_{ij}-1} (1 - r_{ij}) r_{ij}^{n_{ij}} = r_{ij}^{s_{ij}}$, which completes the proof. These results extend the analysis found in Buzacott and Shanthikumar (1993) to systems with multiple facilities. □

Appendix 2. Proof of Lemma 2

If we let Q_i be the number of units of type i that are on order, then $Q_i = Q_{i1} + Q_{i2} + \cdots + Q_{iM}$, and the probability-generating function of the total number of jobs of type i on order is given by

$$g_i(z) = E(z^{Q_i}) = E(z^{Q_{i1}}) E(z^{Q_{i2}}) \cdots E(z^{Q_{iM}}), \tag{A4}$$

which can be rewritten as

$$g_i(z) = \prod_{j=1}^{M} (1 - r_{ij}) \Big/ \prod_{j=1}^{M} (1 - z r_{ij}), \tag{A5}$$

which in turn leads upon a partial fraction expansion to

$$\Pr(Q_i = n_i) = \sum_{j=1}^{M} r_{ij}^{n_i + M - 1} \prod_{k=1}^{M} (1 - r_{ik}) \Big/ \prod_{k \neq j} (r_{ij} - r_{ik}), \tag{A6}$$

provided that $r_{ij} \neq r_{ik}$ for $j \neq k$. In a balanced system with $r_{ij} = r_i$ for $j = 1, \ldots, M$, the total number of units of type i on order has a negative binomial distribution; that is,

$$\Pr(Q_i = n_i) = \binom{n_i + M - 1}{n_i} (1 - r_i)^M r_i^{n_i}. \tag{A7}$$

Let $R_{ij} = \prod_{k \neq j} r_{ij}^{M-1} (1 - r_{ik})/(r_{ij} - r_{ik})$. Then, we can rewrite (A6) as

$$\Pr(Q_i = n_i) = \sum_{j=1}^{M} R_{ij} (1 - r_{ij}) r_{ij}^{n_i}. \tag{A8}$$

From the above, we can obtain expected inventory for each product as

$$\begin{aligned} E(I_i) &= \sum_{n_i=0}^{s_i} (s_i - n_i) \Pr(Q_i = n_i) \\ &= \sum_{j=1}^{M} R_{ij} [s_i - r_{ij}(1 - r_{ij}^{s_i})/(1 - r_{ij})], \end{aligned} \tag{A9}$$

and noting that $s_i = E(I_i) + E(Q_i) - E(B_i)$, the expected number of backorders as

$$E(B_i) = \sum_{j=1}^{M} R_{ij} [r_{ij}^{s_i+1}/(1 - r_{ij})], \tag{A10}$$

in which we use the fact that $E(Q_i) = \sum_{j=1}^{M} r_{ij}/(1 - r_{ij})$. From (A8), we also obtain

$$\Pr(Q_i \geq s_i) = \sum_{j=1}^{M} R_{ij} r_{ij}^{s_i}, \tag{A11}$$

which completes the proof.

Appendix 3. Random Data Generation Procedure

The production rate μ_j, the arrival rates λ_j, and the cost parameters c_{ij} are randomly generated from continuous uniform distributions with ranges [10, 50], [10, 20], and [100, 150]. Unit holding costs are obtained as $h_{ij} = a_{ij} c_{ij}$, where the parameter a_{ij} is also generated randomly from a continuous uniform distribution with range [0.02, 0.30]. Similarly, unit backordering costs are obtained as $b_{ij} = \theta_{ij} h_{ij}/(1 - \theta_{ij})$, where θ_{ij} is randomly generated from a uniform distribution over the range (0.60, 0.99). The parameter θ_{ij} can be viewed as the effective desired fill rate of product i at facility j. To obtain the desired overall utilization ρ, the randomly generated demand rates are rescaled by multiplying each λ_i by $(\sum_{i=1}^{N} \lambda_i / \sum_{i=1}^{N} \mu_i)/\rho$.

Appendix 4. Analysis of Systems with Multiple Classes and Centralized Inventory

The following result characterizes the optimal threshold levels s_i^k, $k \leq K_i$, and the optimal base-stock levels s_i^* for a system with multiple customer classes and centralized inventory as discussed in §7.1.

Lemma A1. *Construct the sequence s_i^k and g_i^k as follows. Set $s_i^0 = g_i^0 = b_{K_i+1} = 0$. Then, for $k \leq K_i$,*

(1) *compute*

$$\hat{r}_{ij}^k = \frac{\alpha_{ij} \sum_{l=1}^{k} \lambda_i^l}{\mu_j - \sum_{l \neq i} \alpha_{lj} \lambda_l} \quad \text{and} \quad R_{ij}^k = (\hat{r}_{ij}^k)^{M-1} \prod_{l \neq j} \frac{(1 - \hat{r}_{il}^k)}{(\hat{r}_{ij}^k - \hat{r}_{il}^k)},$$

(2) *determine s_i^{k+1}, the smallest integer such that $\Delta g_{k+1}(s) > 0$, where*

$$\begin{aligned} \Delta g_i^{k+1}(s) = \sum_{j=1}^{M} R_{ij}^k \big[h_i + b_i^k - (\hat{r}_{ij}^k)^{s - s_i^k} \big[(1 - \hat{r}_{ij}^k)(g_i^k - (h_i + b_i^k) s_i^k) \\ + (h_i + b_i^k) \hat{r}_{ij}^k \big] \big], \quad \text{and} \end{aligned}$$

(3) *evaluate*

$$\begin{aligned} g_i^{k+1} = \sum_{j=1}^{M} R_{ij}^k \Bigg[\left(s_i^k - \frac{\hat{r}_{ij}^k}{1 - \hat{r}_{ij}^k} \right) (h_i + b_i^{k+1}) \\ + \left(g_i^{k+1} - \left(s_i^k - \frac{\hat{r}_{ij}^k}{1 - \hat{r}_{ij}^k} \right) (h_i + b_i^k) \right) (\hat{r}_{ij}^k)^{s - s_i^k} \Bigg]. \end{aligned}$$

The optimal threshold levels are given by s_i^k, for $k \leq K_i$, the optimal base-stock level for product i is equal to $s_i^* = s_i^{K_i+1}$, and the corresponding optimal cost is given by $g_i^{K_i+1}$.

Proof. For any k, denote by g_i^k the optimal average cost when product i has k classes of customers, with backorder costs given by $(b_i^1 - b_i^{k+1}, \ldots, b_i^k - b_i^{k+1})$ and holding costs $h_i + b_i^{k+1}$. (Note that when $k = K_i$, then $b_i^{k+1} = 0$ and we obtain the initial problem.) Denote by $(x_i^0, \ldots, x_i^k)$ the system state with k customer classes, where x_i^0 represents the inventory level and x_i^l, $l \in \{1, \ldots, k\}$, the number of backorders for class i. Also, let X_i^l, $l \in \{1, \ldots, k\}$, denote the corresponding random variables. The instantaneous costs $c_i^k(x_i^0, \ldots, x_i^k)$ are then given by

$$\begin{aligned} c_i^k(x_i^0, \ldots, x_i^k) &= (h_i + b_i^{k+1}) x_i^0 - \sum_{l=1}^{k} (b_i^l - b_i^{k+1}) x_i^l \\ &= c_i^{k-1}(x_i^0, \ldots, x_i^{k-1}) - (b_i^k - b_i^{k+1}) \sum_{l=0}^{k} x_i^l. \end{aligned}$$

Let $Q_i^k = Q_{i1}^k + \cdots + Q_{ij}^k + \cdots + Q_{iM}^k$ be the number of units of type i when there are k classes of customers. The arrival rate for product i at facility j is then equal to $\alpha_{ij}\sum_{l=1}^{k}\lambda_i^l$. Following de Véricourt et al. (2000) we show the result by iterating on the number of customer classes. Assuming that the property is true for k classes of customers for a given base-stock level s, the average cost $g_i^{k+1}(s)$ with $k+1$ customer classes can be shown, after some computations, to be given by

$$g_i^{k+1}(s) = g_i^k \Pr(Q_i^{k+1} \geq s) + (h_i + b_i^{k+1}) \sum_{z=s_i^k+1}^{s} z \Pr(Q_i^{k+1} = s - z) + (b_i^k - b_i^{k+1})(E(Q_i^{k+1}) - s).$$

The function $\Delta g_i^{k+1}(s)$ can then be derived using the probability distribution of Q_i^{k+1} given by Equation (19), with the total arrival rate for product i equal to $\sum_{l=1}^{k}\lambda_i^l$. Because $\Delta g_i^{k+1}(s)$ is convex, s_i^{k+1} is well defined and g_i^{k+1} can be obtained from $g_i^{k+1}(s_i^{k+1})$. □

If we allow stocks to take real values, the previous optimal procedure can be simplified and the total cost for the case of centralized inventory with rationing reduces to

$$z(\boldsymbol{\alpha}, \mathbf{s}^*) = \sum_{i=1}^{N}\left[h_i s_i^* + \sum_{j=1}^{M} c_{ij}\alpha_{ij}\lambda_{ij}\right],$$

where the optimal base-stock levels $s_i^* = s_i^{K_i+1}$ are recursively computed as follows: for $0 \leq k \leq K_i$ we obtain s_i^{k+1} the unique solution of $\sum_{j=1}^{M} R_{ij}^k (\hat{r}_{ij}^k)^{s-s_i^k} = (h_i + b_i^{k+1})/h_i + b_i^k$.

References

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

Bazaraa, M. S., C. M. Shetty. 1979. *Nonlinear Programming.* John Wiley & Sons, Inc., New York.

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. 2004a. On the benefit of pooling in production-inventory systems. Working paper, University of Minnesota, Minneapolis, MN.

Benjaafar, S., M. Elhafsi, F. de Véricourt. 2003. Supplement to demand allocation in multi-product, multi-facility make-to-stock systems. Working paper, University of Minnesota, Minneapolis, MN.

Benjaafar, S., J. S. Kim, N. Vishwanadham. 2004b. On the effect of product variety in a production-inventory system. *Ann. Oper. Res.* **126** 71–101.

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

Cattrysse, D., L. N. Van Wassenhove. 1992. A survey of algorithms for the generalized assignment problem. *Eur. J. Oper. Res.* **60** 260–272.

Fisher, M. L. 1981. The Lagrangian relaxation method for solving integer programming problems. *Management Sci.* **14** 1–18.

Green, L. V., D. Guha. 1995. On the efficiency of imbalance in multi-facility multi-server service systems. *Management Sci.* **41** 179–187.

Gupta, D., S. Benjaafar. 2004. Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis. *IIE Trans.* **36** 529–546.

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

Liu, Z., R. Righter. 1998. Optimal load balancing on distributed homogenous unreliable processors. *Oper. Res.* **46**(4) 563–573.

Martello, S., P. Toth. 1990. *Knapsack Problems: Algorithms and Computer Implementation.* Wiley, New York.

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

Osman, J. H. 1995. Heuristics for the generalized assignment problem: Simulated annealing and tabu search approaches. *OR Spektrum* **17** 211–225.

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

Ross, G. T., R. M. Soland. 1975. A branch and bound algorithm for the generalized assignment problem. *Math. Programming* **8** 91–103.

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

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

de Véricourt, F., Y. Dallery. 2000. Non-optimality of static-priority policies in unreliable two part type production system. *IEEE Trans. Automatic Control* **45** 309–311.

de Véricourt, F., M. Veatch. 2003. Zero-inventory conditions for a two-part type make-to-stock production system. *Queuing System Theory Appl.* **43** 251–266.

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

de Véricourt, F., F. Karaesmen, Y. Dallery. 2001. Assessing the benefits of rationing and scheduling policies for a make-to-stock production system. *Manufacturing Service Oper. Management* **3** 105–121.

de Véricourt, F., F. Karaesmen, Y. Dallery. 2002. Stock allocation for a capacitated supply system. *Management Sci.* **48** 1486–1501.

Wang, Y., R. J. T. Morris. 1985. Load sharing in distributed systems. *IEEE Trans. Comput.* **34** 204–317.

Wein, L. M. 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.

Wolff, R. W. 1989. *Stochastic Modelling and the Theory of Queues.* Prentice-Hall, Engelwood Cliffs, NJ.

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

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

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

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