A Two-Echelon Inventory-Location Problem with Service Considerations

Ho-Yin Mak,[1] Zuo-Jun Max Shen[2]

[1] *Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong*

[2] *Department of Industrial Engineering and Operations Research, University of California, Berkeley, California 94720-1777*

Received 4 January 2008; revised 23 June 2009; accepted 12 July 2009
DOI 10.1002/nav.20376
Published online 7 October 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: We study the problem of designing a two-echelon spare parts inventory system consisting of a central plant and a number of service centers each serving a set of customers with stochastic demand. Processing and storage capacities at both levels of facilities are limited. The manufacturing process is modeled as a queuing system at the plant. The goal is to optimize the base-stock levels at both echelons, the location of service centers, and the allocation of customers to centers simultaneously, subject to service constraints. A mixed integer nonlinear programming model (MINLP) is formulated to minimize the total expected cost of the system. The problem is NP-hard and a Lagrangian heuristic is proposed. We present computational results and discuss the trade-off between cost and service. © 2009 Wiley Periodicals, Inc. Naval Research Logistics 56: 730–744, 2009

Keywords: supply chain design; service parts inventory systems; response time requirements

1. INTRODUCTION

Logistics is a vital aspect of business management. Logistics (transportation, inventory, and administrative) costs amounted to 8.7% of the US GDP in 2002 [22]. This proportion decreased significantly from the early 1980s (16.2% in 1981), suggesting an improvement in efficiency of logistics management over the past two decades. Traditionally, companies have managed distribution and storage decisions separately, in part due to the complexity of combining them. Furthermore, the tactical and operational decisions of stocking and distribution are considered separately from strategic decisions of facility location and network design. Shen et al. [9], Daskin et al. [10] and Candas and Kutanoglu [7] have shown that ignoring the effect of inventory in facility location decisions can lead to suboptimal supply chain designs. However, it remains a challenging task to integrate facility location with inventory management and distribution decisions, especially with considerations of customer service levels.

In this article, we are concerned with the problem of designing a two-echelon service parts supply chain under customer service level constraints. This problem is of importance because spare parts inventory is expensive. As pointed out by IBM [18], 30–60% spare parts inventory reduction at 30 client sites can generate a business value of US$ 10–500 million per year. About a decade ago, inventory investment accounted for about 28% of the total cost of service parts logistics systems, whereas warehousing and "other" (including general administrative, personnel and miscellaneous) costs accounted for about 35% and transportation accounted for only 8.4 % [9]. Although the absolute monetary values of these costs have changed over the years, the ratios between these costs should remain at similar orders of magnitude. Because they can amount up to one-third of the total cost, inventory investments must be included in consideration when designing service parts systems.

In our problem setting, there is a central manufacturing plant with limited production and storage capacity. The firm locates service centers (SCs) that hold inventory to satisfy demand from spatially dispersed customers and replenish inventory by ordering from the plant. All SCs and the plant manage inventory using continuous review base stock $(S-1, S)$ policies. The problem is to determine the number and location of DCs, the assignment of customers to SCs and the base stock levels at SCs and the plant, subject to response time requirements.

Base stock $(S-1, S)$ inventory policies are suitable for products with relatively low demand and high inventory holding costs. Moinzadeh and Lee [23] provide an analytical model to determine whether base stock policies are optimal given certain problem parameters. Their results suggest that

Correspondence to: Z.-J.M. Shen (shen@ieor.berkeley.edu)

a base stock policy is optimal when demand rates are low and setup costs are low relative to holding cost rates. This class of policies is used in spare parts inventory systems, in which a failed part is replaced by a new one from inventory. If the failure rate is low, a base stock policy is desirable. In this article, we assume that the plant and SCs manage inventory using a base stock policy. Such systems exist in a wide range of sectors, including military, airlines, and computer manufacturing. We use a spare parts inventory system as our motivating example, although the model is applicable to a variety of production-inventory systems as well.

One might think that when demand rates are low, it would always be better to ship directly from the plant to the customer instead of keeping inventory at SCs. However, in many applications (especially in spare parts systems), the customers are sensitive to response times. It is then desirable to store inventory at SCs that are close to customers located far from the plant. Caglar et al. [6] state that such a system is suitable for spare parts systems where SCs equipped with spare parts inventory and technicians are located close to customers. For example, In IBM's multi-echelon service parts system, items are sent to customers from nearby stocking locations when needed [8, 20]. In this article, we consider this class of inventory systems, in which customers wish to be served within a time limit and thus SCs must be located.

The remainder of this article is organized as follows: In Section 2, we review the related literature on facility location and multi-echelon inventory models. Then we present the formulation of the model, its properties, and the solution approach in Sections 3 and 4, respectively. Finally, we present results of computational experiments in Section 5 and conclude in Section 6.

2. LITERATURE REVIEW

Location analysis has been studied extensively in the operations research, economics, and geography literature. Daskin [10], Drezner [12], and Drezner and Hamacher [13] provide excellent reviews of the theory. In these models, we are concerned with the strategic issue of selecting candidate sites to locate facilities and cover demand. However, inventory considerations have traditionally not been included in the facility location literature.

One of the earliest and most influential articles in the area of multi-echelon inventory management is by Sherbrooke [32]. For a two-echelon inventory system for repairable parts operating under $(S-1, S)$ policies, he develops the METRIC model to approximate expected backorders at each facility. Another classical article is written by Graves [17] who proposes a two-parameter approximation as an alternative to METRIC. He shows numerically that the two-parameter approximation is more accurate than METRIC. For textbook treatments of multi-echelon inventory theory including the above models, see, for example, [1, 32].

There has been research on two-echelon inventory systems that consider time-based service requirements. Kutanoglu [20] studies a system with the possibility of emergency lateral transshipments between the local facilities (i.e., SCs). His evaluation model provides important insights including how time-based service requirements are more relevant in service parts systems than fill rates, and how emergency lateral transshipment improves response time performances. Caglar [5] and Caglar et al. [6] develop algorithms to optimize stocking levels at both levels under response time constraints.

Several articles study inventory control under time-sensitive service requirements in different supply chain settings. Lee and Billington [20], motivated by the operations of HP's Deskjet Printer supply chain, study the performance measures of fill rate, mean delay, and variance of delay, in a general material flow network. Ettl et al. [15] formulate an optimization model which takes into account actual delays due to stock-outs. More recently, Simchi-Levi and Zhao [3, 5] study a tree network supply chain facing more realistic "transit times" instead of i.i.d. stochastic lead times. They also provide algorithms to minimize inventory costs subject to service requirements requiring that the delivery lead time be shorter than a specified threshold with a certain high probability (i.e., the service level).

Although the articles mentioned earlier provide important insights on the tactical and operational aspects of two-echelon inventory systems, the focus of our article is on the strategic design phase.

Shen et al. [29] propose an *integrated* approach of supply chain design. By explicitly including inventory cost terms in the strategic facility location model, the supply chain structure is optimized under an objective function that much better reflects the true operating cost than does the traditional distance-based objective. They show that the integrated approach produces better supply chain designs than the traditional sequential design-operations optimization approach. Building on the integrated framework, various articles have been published. Daskin et al. [11] and Shu et al. [34] improve the solution technique. Shen and Qi [31] further incorporate operational routing costs into the model. Shen and Daskin [30] evaluate the trade-offs between cost minimization and service maximization. Ozsen *et al.* [25, 26] consider the effect of storage capacity at facilities under single and multiple sourcing. For more information regarding these models, see Shen [28] which provides a more detailed survey. Although the models point out the importance of the integrated planning approach, only single-echelon problems in which inventory is only held at one level of facilities have been studied.

Multi-echelon inventory-location models have received relatively little attention. Nozick and Turnquist [24] consider the location of distribution centers (lower echelon) in a two-echelon system with inventory held at distribution centers and a plant operating under $(S-1, S)$ policies. They utilize a linear approximation for safety stock costs as a function of the

number of distribution centers. The resulting location model is structurally identical to the uncapacitated fixed charge problem, with inventory costs included in the fixed location cost as a constant.

More recently, Candas and Kutanoglu [7] consider a multi-commodity two-echelon inventory-location problem in which stocking levels and fill rates can be optimized to achieve a system-wide time-based service level. They formulate a non-linear integer programming model and propose a linearization-based technique to solve small to medium sized instances. Their results show that the integrated approach can yield solutions that achieve the same service level at a lower cost.

Despite the important contributions of the articles mentioned earlier, all of them treat replenishment lead time as deterministic with the exception of the article by Nozick and Turnquist which treats the replenishment process as an M/G/∞ queue. Only a small number of inventory-location models in the literature allow replenishment lead times to be stochastic. Eskigun et al. [17], Eskigun et al. [15], and Sourirajan et al. [36] study different supply chain network design problems with stochastic lead times. Finally, Benjaafar et al. [2] study the problem of locating facilities and managing inventory in a single-echelon system. All of the articles discussed in this paragraph utilize approximate or exact queuing formulas to model congestion in stochastic lead times.

In this article, we build on the idea of Benjaafar et al. and consider a more general two-echelon system. Our problem also includes consideration of response time and distance constraints. To the best of our knowledge, there is no published work on joint inventory-location problems that considers both stochastic replenishment lead time and response time requirements.

3. MODEL FORMULATION

3.1. Basic Formulation

We consider a single-product supply chain which consists of a single plant (upper echelon), a set of SCs (lower echelon) and a set of customers. The plant manufactures the item and holds inventory to meet demands from SCs. The SCs hold inventory to fulfill customer orders. Each customer places orders at an assigned SC following a Poisson process, consistent with the classical exponential failure model in reliability theory. We assume that orders from different customers are independent. This assumption is valid when customers do not interact with each other. We assume that the plant and the SCs each have fixed amount of storage space to hold inventory and operate with continuous review $(S-1, S)$ replenishment policies. As the items are typically expensive in many spare parts applications, it is also possible to interpret the capacity restriction as a budget constraint.

When a customer places an order, the SC sends one unit of the product from its inventory (if there is stock) to the customer and immediately places an order with the plant. When the plant receives an order, it sends one unit of the product from its inventory (if there is stock) to the SC, and immediately releases an order of one unit to its production line. If a service center or the plant is out of stock, the demand is backordered until they are filled. Backorders are handled in a first-come, first-served (FCFS) manner. Finished goods from the production line are used to fill backorders or are stored as inventory at the plant immediately after they leave the production line.

The problem is to determine simultaneously the optimal number and location of SCs, the assignment of customers to the opened SCs and the inventory stocking levels at SCs and the plant. The costs considered in the model include fixed location costs of the SCs, transportation costs from the plant to the SCs and from the SCs to the customers, and inventory holding costs at the plant and at the SCs. We begin by introducing the following notation:

3.1.1. Sets

I = Set of customers
J = Set of candidate service center locations

3.1.2. Cost Parameters

f_j = Fixed cost of opening a SC at location j
h_j = Holding cost per unit of inventory per unit time at location j
p = Backorder cost per unit of inventory per unit time
d_{ij} = Shipping cost per unit from location j to customer i

3.1.3. Demand and Other Parameters

λ_i = Demand rate (Poisson) at customer i
λ = Total demand of all customers $(= \sum_{i \in I} \lambda_i)$
μ = Order processing or manufacturing rate at the plant
ρ = Utilization rate of the plant $(= \lambda/\mu)$
τ = Average response time requirement
α_j = Deterministic transportation lead time from the plant to location j
$d_{\max}$ = Maximum distance allowed between customer and the assigned SC
a_{ij} = 1 if customer i is within $d_{\max}$ distance from candidate location j, 0 otherwise
C_j = Storage space available at candidate SC location j, in number of units of the product

3.1.4. Decision Variables

X_j = 1 if a SC is located at j, 0 otherwise
Y_{ij} = 1 if demand at customer i is assigned to SC at j, 0 otherwise

S_j = Base stock level maintained at SC at j
S_0 = Base stock level maintained at the plant

3.1.5. Service Variables

$\bar{I}_j$ = Steady state expected inventory level at SC j
$\bar{B}_j$ = Steady state expected backorder level at SC j
$\bar{W}_j$ = Steady state expected response time at SC j
$\bar{I}_0$ = Steady state expected inventory level at the plant
$\bar{B}_0$ = Steady state expected backorder level at the plant
$N_j(t)$ = Number of replenishment orders made by DC j that has not yet arrived by time t

Using the above notation, the model can be formulated as follows:

$$\min \sum_{j\in J}\left(f_j X_j + h_j \bar{I}_j + p\bar{B}_j + \sum_{i\in I} d_{ij}\lambda_i Y_{ij}\right) + h_0 \bar{I}_0 \tag{1}$$

Subject to:

$$\sum_{j\in J} Y_{ij} = 1, \quad \text{for each } i \in I \tag{2}$$

$$Y_{ij} \le a_{ij} X_j, \quad \text{for each } i \in I, j \in J \tag{3}$$

$$S_j \le C_j X_j, \quad \text{for each } j \in \{0\} \cup J \tag{4}$$

$$\bar{W}_j \le \tau, \quad \text{for each } j \in J \tag{5}$$

$$X_j \in \{0,1\}, \quad \text{for each } j \in J \tag{6}$$

$$Y_{ij} \in \{0,1\}, \quad \text{for each } i \in I, j \in J \tag{7}$$

$$S_j \ge 0, \text{ integer}, \quad \text{for each } j \in \{0\} \cup J \tag{8}$$

The objective (1) is to minimize the sum of the (annualized) fixed location costs, shipment costs, inventory holding costs at the plant and the SCs, and backorder costs at SCs. Backorders at the plant do not incur a monetary cost as they are considered "internal" to the system. Constraints (2) require that all customers must be assigned to SCs. Constraints (3) state that customer assignments cannot be made to a candidate location unless a SC is opened and that the resulting distance is shorter than $d_{\max}$. Constraints (4) require that the base stock level at a SC cannot exceed the storage capacity. The service time constraints (5) state that the expected response time (time between order arrival and shipment made) cannot exceed the required level. Finally, (6–8) are nonnegativity and integrality constraints on the decision variables.

In this model, the response requirements are that (i) the distance between any customer and the assigned SC is no longer than a specified limit $d_{\max}$, which is captured by constraint (3); and (ii) the average time between receiving an order and sending out the item at any SC must not exceed the service guarantee (5). Models in the spare parts inventory management literature typically consider either shortage costs or service requirements. The models that consider service requirements have such constraints either on service level [8] or response time [5]. The former type of constraints limit the chance of stocking out and essentially ignore the differences between stock-outs with short response times (e.g., an hour) and those with long response times (e.g., 2 days). In our two-echelon system, stock-outs at SCs may have a short response time if the plant has the item in stock and ships immediately or a long lead time if the plant also has a stock-out. As we are mainly concerned with designing spare parts inventory systems, a response time requirement is more appropriate.

Considering each SC as a queuing system, it is possible to apply Little's law and replace (5) by the following:

$$\bar{B}_j \le \tau \sum_{i\in I} \lambda_i Y_{ij} \tag{9}$$

Before proposing the solution algorithm, we would like to express the inventory and backorder levels in our formulation in terms of the decision variables, i.e., (X, Y, S).

3.2. Inventory Level at the Plant

Under the $(S-1, S)$ policy, the demand faced by the plant is the superposition of the demand processes at the customers, which are independent and Poisson. Therefore, the production line at the plant behaves as a queue with Markovian arrivals. Let N_0 denote the steady state number of orders in the queuing system (in line and in service). It is then standard to express:

$$\bar{I}_0 = S_0 - E[N_0] + \bar{B}_0 \tag{10}$$

$$\bar{B}_0 = E[N_0] - \sum_{s=0}^{S_0-1}[1 - F_0(s)] \tag{11}$$

$$\text{where } F_0(s) = \sum_{m=0}^{s} P(N_0 = m)$$

By substituting steady state probabilities into the above formulas, we can easily obtain the expected plant inventory and backorder levels for different manufacturing queue structures. For example, we may substitute the standard formula for the steady state distribution of the number in system of the $M/M/c$ queue. It is also possible to allow a batch ordering policy at the plant.[1] Suppose the plant sends a job request to the production line after receiving $Q > 0$ orders. Then

[1] We would like to thank an anonymous referee for pointing out this fact.

the interarrival time distribution at the manufacturing queue is Erlang with parameters (Q, λ). Then we may substitute the steady state distribution of the $E_Q/M/1$ queue as given in [19], for example.

For simplicity we will only consider the M/M/1 case. Buzacott and Shanthikumar [4] show that the average inventory and backorder levels at the plant are given by:

$$\bar{I}_0 = S_0 - \frac{\rho}{1-\rho}(1-\rho^{S_0}) \tag{12}$$

$$\bar{B}_0 = \frac{\rho^{S_0+1}}{1-\rho} \tag{13}$$

$$\bar{W}_0 = \frac{\bar{B}_0}{\lambda} = \frac{\rho^{S_0+1}}{\lambda(1-\rho)} \tag{14}$$

3.3. Inventory Level at SCs

Similar to the case for the plant, the steady state expected inventory and backorder levels at each SC as follows:

$$\bar{I}_j = S_j - E[N_j] + \bar{B}_j \tag{15}$$

$$\bar{B}_j = E[N_j] - \sum_{s=0}^{S_j-1}[1-F_j(s)] \tag{16}$$

$$\text{where} F_j(s) = \sum_{m=0}^{s} P(N_j = m)$$

By disaggregating the backorders at the plant due to each SC [see [17], for example], the expected value of N_j is given by:

$$E[N_j] = \frac{\lambda_j}{\lambda}\bar{B}_0 + \lambda_j\alpha_j = \frac{\lambda_j}{\lambda}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_j\alpha_j \tag{17}$$

The exact algorithm proposed by Graves [17] to obtain the distribution of N_j is too computationally costly to be included as a subroutine of a complex optimization problem like ours. Therefore, we construct a METRIC-like method by approximating the distribution of N_j with a Poisson distribution with mean given in (17). The METRIC-like approximation makes the expected SC inventory levels a convex function of assigned demand (Lemma 1). This property can be exploited to design efficient solution algorithms. Although in principle one can also use the negative binomial approximation suggested by Graves [17], doing so requires rounding because the parameter r has to be integer-valued. This makes the resulting expected inventory and backorder level expressions nondifferentiable with respect to assigned demand and increases the difficulty of optimization. Therefore, we proceed with the reasonably accurate and much more tractable METRIC-like approximation.

Using the METRIC-like method, we approximate N_j in (12) and (13) with a Poisson random variable. Therefore, we replace $F_j(s)$ by the Poisson CDF with mean $\lambda_j\bar{L}_j$, where λ_j is the demand assigned to SC j $(= \sum_{i\in I}\lambda_i Y_{ij})$. $\bar{L}_j$ is the expected replenishment lead time which consists of expected response time of the plant and the delivery lead time:

$$\bar{L}_j = \bar{W}_0 + \alpha_j = \frac{\rho^{S_0+1}}{\lambda(1-\rho)} + \alpha_j \tag{18}$$

With the manipulations described earlier, it is possible to reformulate the problem as follows:

$$\min \sum_{j\in J}\left\{ f_j X_j - pS_j + (h_j+p)\sum_{s=0}^{S_j-1}F_j(s) + \sum_{i\in I}\left[p\left(\frac{\rho^{S_0+1}}{\lambda(1-\rho)} + \alpha_j\right) + d_{ij}\right]\lambda_i Y_{ij}\right\} + h_0\left[S_0 - \frac{\rho}{1-\rho}(1-\rho^{S_0})\right] \tag{19}$$

Subject to:

$$\sum_{j\in J} Y_{ij} = 1, \quad \text{for each } i \in I \tag{20}$$

$$Y_{ij} \le a_{ij}X_j, \quad \text{for each } i \in I, j \in J \tag{21}$$

$$S_j \le C_j X_j, \quad \text{for each } j \in \{0\}\cup J \tag{22}$$

$$\left[\frac{\rho^{S_0+1}}{\lambda(1-\rho)} + \alpha_j - \tau\right]\sum_{i\in I}\lambda_i Y_{ij} \le \sum_{s=0}^{S_j-1}[1-F_j(s)], \quad \text{for each } j \in J \tag{23}$$

$$X_j \in \{0,1\}, \quad \text{for each } j \in J \tag{24}$$

$$Y_{ij} \in \{0,1\}, \quad \text{for each } i \in I, j \in J \tag{25}$$

$$S_j \ge 0, \text{ integer}, \quad \text{for each } j \in \{0\}\cup J \tag{26}$$

In the next section, we present the solution approach based on Lagrangian relaxation.

4. SOLUTION APPROACH

4.1. Obtaining a Lower Bound

We begin by proving the following model property that provides an upper bound on the maximum plant base stock level in many instances.

PROPERTY 1: If $p = 0$ and $\tau > \alpha_{max} = \max_{j\in J}\{\alpha_j\}$, i.e., there is no backorder cost and the service time requirement is longer than the pure transportation lead time between

the plant and any candidate SC site, then an upper bound to the plant base stock level S_0 exists. This is given by:

$$S_0^{\max} = \min\left\{S \geq 0 : \frac{\rho^{S+1}}{\lambda(1-\rho)} + \alpha_{\max} \leq \tau\right\} \tag{27}$$

PROOF: The definition (27) and the condition $\tau > \alpha_{\max}$ imply that for any values of Y_{ij} and S_j, the left hand side of (23) will be nonpositive. As the right hand side is nonnegative, this constraint always holds when $S_0 = S_0^{\max}$. S_0 is only constrained by the response time constraints and nonnegativity, therefore $S_0^{\max}$ is feasible for all values of the other decision variables. As the objective function (19) is strictly increasing in S_0, any solution with $S_0 > S_0^{\max}$ will be suboptimal. ■

The above property states that if there is no backorder cost and all the deterministic transportation lead times between the SCs and the plant are less than the response time requirement, there is a plant base stock level that ensures all response time constraints are satisfied with all SCs holding no inventory. The optimal plant base-stock level will not exceed this level.

The property, in addition to the existence of a capacity constraint and the assumption that overall system demand is low, implies that the range of stock levels that we need to consider to satisfy the response time constraints is small and is bounded above by the capacity. Similar properties have likewise been exploited by Candas and Kutanoglu [7] in developing solution algorithms. For our problem, it is possible to solve a small number of continuous problems by fixing the plant base-stock level to each of the possible values. The best (in terms of cost) of these continuous solutions will be the optimal solution for the original problem.

When the plant base stock level is fixed, all terms that depend only on S_0 become constants. For instance, the plant holding cost will be a constant and can be ignored when solving the restricted problem for the current value of S_0. We choose to relax the assignment constraints (20) and the service constraints (23) in the restricted problem. Given dual multipliers π_i and θ_j corresponding to constraints (20) and (23), respectively, the Lagrangian problem decomposes by candidate SC locations into subproblems of the following form (for each $j \in J$):

$$\min_{Y_{ij}, S_j} (h_j + p + \theta_j) \sum_{s=0}^{S_j - 1} F_j(s) - (p + \theta_j) S_j + \sum_{i \in I} [(d_{ij} + p\alpha_j + \theta_j \alpha_j - \theta\tau)\lambda_i - \pi_i] Y_{ij} + \frac{\theta_j \rho^{S_0+1}}{\lambda(1-\rho)} \sum_{i \in I} \lambda_i Y_{ij}$$

Subject to:

$$S_j \leq C_j \tag{28}$$

$$Y_{ij} \leq a_{ij}, \quad \text{for each } i \in I \tag{29}$$

$$Y_{ij} \in \{0, 1\}, \quad \text{for each } i \in I \tag{30}$$

$$S_j \geq 0, \quad \text{integer} \tag{31}$$

To solve the above subproblem, we use the fact that the SCs have limited storage capacity. In practice, the possible range of S_j is small even in the absence of a physical storage limit because the management may want to limit the amount of inventory held due to the high item cost. Therefore, a promising approach is to solve the subproblem by fixing S_j to each of the possible values. Moreover, we relax the integrality of Y_{ij}, allowing them to take on any value between 0 and 1. The continuous relaxation provides a lower bound for the Lagrangian relaxation approach. As shown in Section 6, the optimality gap of the Lagrangian algorithm is small in general, suggesting that the continuous relaxation of the subproblem is tight. We will now show that the continuous relaxation of the subproblem can be solved efficiently.

With the value of S_j fixed, the continuous subproblem can be expressed as:

$$\min_{0 \leq Y_{ij} \leq a_{ij}} \quad \sum_{i \in I} A_{ij} Y_{ij} + G\left(\sum_{i \in I} \lambda_i Y_{ij}\right) \tag{32}$$

$$\text{where} \quad G(x) = (h_j + p + \theta_j) \sum_{s=0}^{S_j - 1} F_j(s, x)\ ,\ F_j(s, x) = \sum_{m=0}^{s} \frac{e^{-x\bar{L}_j}(x\bar{L}_j)^m}{m!} \tag{33}$$

$$\text{and} \quad A_{ij} = (d_{ij} + p\alpha_j + \theta_j\alpha_j - \theta_j\tau)\lambda_i - \pi_i + \frac{(\theta_j + p)\lambda_i \rho^{S_0+1}}{\lambda(1-\rho)} \tag{34}$$

The solvability of the subproblem depends on properties of $G(x)$ as discussed later.

LEMMA 1: The function $G(x)$ is decreasing and convex in x when $x \geq 0$.

PROOF: Note that $G(x)$ represents the holding and penalty costs on the steady state expected inventory level defined in (15,16), with demand rate x assigned to the SC with base-stock level S_j. Taking the partial derivative of $F_j(s)$ with respect to the assigned demand x (dropping subscript j for simplicity):

$$\frac{\partial}{\partial x}F(s,x) = -Le^{-x\bar{L}} + \sum_{m=1}^{s}\frac{\bar{L}^m}{m!}\left[mx^{m-1}e^{-x\bar{L}} - \bar{L}x^m e^{-x\bar{L}}\right]$$
$$= -\frac{L^{s+1}}{s!}x^s e^{-x\bar{L}} \le 0$$
$$\frac{\partial}{\partial x}G(x) = (h+p+\theta)\sum_{s=0}^{S-1}\frac{\partial}{\partial x}F(s,x) \le 0 \quad (35)$$

Therefore, $G(x)$ is decreasing in x. Taking second derivative of $F(s)$:

$$\frac{\partial^2}{\partial x^2}F(s) = \begin{cases} \bar{L}^2 e^{-x\bar{L}} & t=0 \\ -\dfrac{\bar{L}^{s+1}x^{s-1}e^{-x\bar{L}}}{(s-1)!} + \dfrac{\bar{L}^{s+2}x^s e^{-x\bar{L}}}{s!} & \text{otherwise} \end{cases}$$
$$\frac{\partial^2}{\partial x^2}G(x) = (h+p+\theta)\sum_{s=0}^{S-1}\frac{\partial^2}{\partial x^2}F(s)$$
$$= (h+p+\theta)\frac{\bar{L}^{S+1}x^{S-1}e^{-x\bar{L}}}{(S-1)!} \ge 0 \quad (36)$$

Therefore $G(x)$ is convex. ∎

COROLLARY 1: The continuous subproblem (32) is a convex optimization problem.

Corollary (1) suggests that it is possible to solve the continuous subproblem using standard convex optimization solvers. Alternatively, we propose an efficient procedure that exploits the problem properties. The algorithm is based on the following theorems, similar to the results of Ozsen et al. [25]. To begin, we divide customers within the coverage distance (i.e., $a_{ij} = 1$) into two subsets I_j^+ and I_j^-, with $A_{ij} > 0$ and $A_{ij} \le 0$, respectively. Suppose there are m customers in the subset I_j^+ which are sorted in the following order:

$$0 \le \frac{A_{1j}}{\lambda_1} \le \frac{A_{2j}}{\lambda_2} \le \cdots \frac{A_{mj}}{\lambda_m} \quad (37)$$

THEOREM 1: There exists an optimal solution Y_j^* to the continuous subproblem (32) where:

1. $Y_{ij} = 1$ for all $i \in I_j^-$.
2. At most one of the assignment variables Y_{ij}^* takes on a (strictly) fractional value.
3. If $Y_{kj}^* > 0$ for some $1 \le k \le m$, then $Y_{ij}^* = 1$ for all $1 \le i \le k-1$.
4. $Y_{ij} = 0$ for all i where $a_{ij} = 0$.

PROOF: The first property follows from Lemma 1 which states that $G(x)$ is decreasing. Suppose that in an optimal solution, $Y_{ij} < 1$ for some $i \in I_j^-$. Then by adding a positive quantity $\epsilon \le 1 - Y_{ij}$, the objective function value decreases, as the linear term does not increase ($A_{ij} \le 0$) and the nonlinear term decreases with the argument of $G(.)$ increasing. Therefore there is a contradiction.

To prove the second property, let Y_j' be an optimal solution vector to the continuous subproblem with Y_{kj}' and Y_{lj}' taking on strictly fractional values, $k, l \in I_j^+$ and $k < l$. Let the objective value of this solution be Z'. Then define a new solution Y_j'' as follows:

$$Y_{ij}'' = \begin{cases} Y_{ij}' & \text{if } i \ne l,k \\ Y_{kj}' + \epsilon & \text{if } i = l \\ Y_{lj}' - \frac{\lambda_k}{\lambda_l}\epsilon & \text{if } i = l \end{cases} \quad (38)$$

Let $\epsilon = \min\{1 - Y_{kj}', \frac{\lambda_l}{\lambda_k}Y_{lj}'\}$ which implies that Y_j'' is feasible. Denote the objective value of the solution with Y_j'' by Z''. Then the difference between the two solutions is given by:

$$Z'' - Z' = \epsilon\left(A_{kj} - A_{lj}\frac{\lambda_k}{\lambda_l}\right) + G\left(\sum_{i\in I}\lambda_i Y_{ij}' + \lambda_k\epsilon - \epsilon\lambda_l\frac{\lambda_k}{\lambda_l}\right) - G\left(\sum_{i\in I}\lambda_i Y_{ij}'\right)$$
$$= \epsilon\left(A_{kj} - \frac{A_{lj}}{\lambda_l}\lambda_k\right)$$
$$\le \epsilon\left(A_{kj} - \frac{A_{kj}}{\lambda_k}\lambda_k\right) = 0 \quad (39)$$

The inequality (39) holds because k is ranked before l in I_j^+. The above implies that Y_j'' is optimal. Recall that $\epsilon \le \min\{1 - Y_{kj}', \frac{\lambda_l}{\lambda_k}Y_{lj}'\}$. For each of the possible values:

1. If $\epsilon = 1 - Y_{kj}'$, $Y_{kj}'' = 1, 0 < Y_{lj}'' < 1$
2. If $\epsilon = \frac{\lambda_l}{\lambda_k}Y_{lj}'$, $Y_{lj}'' = 0, 0 < Y_{kj}'' < 1$

In both cases, the number of variables with strictly fractional values is reduced by one without increasing the objective value. We may repeat this process until there is only one fractional value in the solution, which proves the second property.

The third property can be proven using a similar interchange argument. ∎

The above theorem suggests that there is at most one assignment variable with a fractional value at optimality. The next theorem allows us to find such a value.

THEOREM 2: Consider a solution satisfying the properties described in Theorem 1 in which there is exactly one i^*

where $1 \le i^* \le m$ where $0 < Y_{i^*j} < 1$. Then the solution is optimal if:

$$\frac{A_{i^*j}}{\lambda_{i^*}} = (h_j + p + \theta_j) \sum_{s=0}^{S_j-1} \frac{\bar{L}^{s+1}}{s!} x^s e^{-x\bar{L}} \tag{40}$$

where $x = \sum_{i \in I} \lambda_i Y_{ij}$

PROOF: The Karush-Kuhn-Tucker (KKT) optimality conditions for the continuous subproblem are given by:

Feasibility Conditions

$$0 \le Y_{ij} \le 1, \quad \text{for each } i \in I \tag{41}$$
$$u_i, v_i \ge 0, \quad \text{for each } i \in I \tag{42}$$

Gradient Conditions:

$$A_{ij} + (h_j + p + \theta_j) \sum_{s=0}^{S_j-1} \frac{\partial}{\partial Y_{ij}} F_j(s) - u_i + v_i = 0, \quad \text{for each } i \in I \tag{43}$$

Complementary Slackness Conditions

$$u_i Y_{ij} = 0, \quad \text{for each } i \in I \tag{44}$$
$$v_i (Y_{ij} - 1) = 0, \quad \text{for each } i \in I \tag{45}$$

If Y_{ij} takes on a fractional value, then $u_i = v_i = 0$ by (44,45). Then (43) becomes (40). The continuous subproblem is convex by Corollary 1, therefore the KKT solution as described earlier is a global minimum. ■

Based on the above results, we propose the following algorithm to solve subproblem (32) for a given j with S_j fixed:

ALGORITHM 1:

Step 1: Partition the set of customers I into:

$$I_j^+ = i \in I : A_{ij} > 0$$
$$I_j^- = i \in I : A_{ij} \le 0$$

Step 2: Sort customers in the subset I_j^+ such that:

$$\frac{A_1 j}{\lambda_1} \le \frac{A_2 j}{\lambda_2} \le \cdots \frac{A_m j}{\lambda_m}$$
$$\text{where } m = |I_j^+|$$

Step 3: Fix $Y_{ij} = 1$ for all $i \in I_j^-$.

Step 4: Compute the following:

$$\hat{A}_{0j} := \sum_{i \in I_j^-} A_{ij}$$
$$\hat{D}_{0j} := \sum_{i \in I_j^-} \lambda_i$$

Define:

$$R_j(i, x) := \frac{A_{ij}}{\lambda_i} - (h_j + p + \theta_j) \sum_{s=0}^{S_j-1} \frac{\bar{L}^{s+1}}{s!} x^s e^{-x\bar{L}} \tag{46}$$

Let $k = 1$ and go to Step 5.

Step 5: Compute the following:

$$\hat{A}_{kj} := \hat{A}_{k-1,j} + A_{kj}$$
$$\hat{D}_{kj} := \hat{D}_{k-1,j} + \lambda_k$$

Step 5a: If $R_j(k, D_{k-1}) < 0$ and $R_j(k, D_k) > 0$, a root to equation (40) exists in the interval $D_{k-1,j} < x < D_{kj}$. This root corresponds to an optimal solution with $0 < Y_{kj} < 1$, $Y_{ij} = 1$ for $i = 1 \ldots k-1$ and $\sum_i \lambda_i Y_{ij} = x$. Apply a bracketed root-finding algorithm such as Brent's method (see, for example, Chapter 4 of ref. [3]) to find the root to the equation $R_j(k, x) = 0$, given by x^*. Then the algorithm terminates and the optimal solution is given by:

$$Y_{ij}^* = \begin{cases} 1 & \text{if } i \in I_j^- \cup \{1 \ldots k-1\} \\ \dfrac{x^* - D_{i-1,j}}{\lambda_i} & i = k \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

Step 5b: If $R_j(k, D_{k-1})$ and $R_j(k, D_k)$ have the same sign, compute:

$$\hat{P}_{kj} := \hat{A}_{kj} + G(\hat{D_{kj}})$$

Increment k by 1. If $k \le m$, go to the beginning of Step 5. If $k > m$, go to Step 6.

Step 6: If we loop through the set I_j^+ once without finding a fractional solution, the optimal cost is given by $\hat{P}_{k^*j} = \min_{k=1\ldots m} \hat{P}_{kj}$. The optimal solution is given by:

$$Y_{ij}^* = \begin{cases} 1 & \text{if } i \in I_j^- \cup \{1 \ldots k^*\} \\ 0 & \text{otherwise} \end{cases} \tag{48}$$

COROLLARY 2: As the algorithm proceeds, if for any given $i^* \in I_j^+$ and S_j we find that the LHS of (40) is greater than the RHS, there will be no solution to (40) for any $i \ge i^*$.

PROOF: The LHS of (40) is increasing in i. The RHS is decreases with added demand. Therefore, if the LHS

is greater than the RHS, there cannot be a solution for $i \geq i^*$. ■

The above corollary allows us to save computation time as it specifies conditions under which the existence of a root to (40) does not need be checked.

The most costly step computationally in the algorithm is solving for the root of equation (40). The algorithm checks the necessary condition for a root to exist within the interval in question. In our computational tests, we observe that a root seldom exists. This suggests that there often exists an integral optimal solution to the subproblem (32). In the vast majority of iterations, the root-finding procedure is not called.

The next most costly step in the algorithm is sorting the set I_j^+, which has complexity $|I_j^+| \log |I_j^+|$. However, the set needs to be sorted only once to solve the subproblems for all possible values of S_j.

4.2. Obtaining an Upper Bound

Section 4.1 outlines a method to find a lower bound for the problem given a set of Lagrangian multipliers. In each iteration of the Lagrangian procedure, we make use of the lower bound solution to find a upper bound (feasible) solution to the problem. Recall that the service constraints (23) and customer assignment constraints (20) are relaxed. Therefore, a lower bound solution represents a system in which these constraints may be violated. We now outline a heuristic procedure to construct a feasible solution using the lower bound solution.

We begin by opening the SCs that are opened in the lower bound solution and setting the plant base stock level equal to the value obtained in the lower bound solution. Then, we partition the set of customers into three groups and determine their assignments in order as follows:

1. The ones that are assigned to exactly one SC in the lower bound solution;
2. The ones that are assigned to more than one SC in the lower bound solution; and
3. The ones that are unassigned in the lower bound solution.

We first try to assign each customer to the nearest SC that it is assigned to in the lower bound solution. If doing so is infeasible (i.e., increases base stock level above capacity), we assign the customer to the open SC that results in the smallest cost increase. For each feasible assignment, the optimal base stock levels at SCs and thus inventory-related costs can be computed easily by the observation that the optimal base-stock level at a SC is the minimum of the newsvendor solution and the upper bound allowed by storage capacity.

4.3. Summary of Lagrangian Relaxation Algorithm

We summarize the algorithm discussed later:

ALGORITHM 2:

Step 0: Iteration count $n = 1$. Initialize dual multipliers (θ^1, π^1). They can be set at any feasible value (e.g., all equal to 0).

Step 1: Find lower bound using the procedure described in Section 4.1 by solving subproblem (32) for each j and $S_j = 0 \ldots C_j$ using Algorithm 1. Let the resulting lower bound value be LB^n and the lower bound solution be $(\underline{X}^n, \underline{Y}^n, \underline{S}^n)$.

Step 2: Find an upper bound with the procedure described in Section 4.2, using lower bound solution $(\underline{X}^n, \underline{Y}^n, \underline{S}^n)$ as input. If this upper bound is smaller than the current best upper bound, let its cost be UB and the corresponding (feasible) solution be $(\bar{X}, \bar{Y}, \bar{S})$. If $\frac{UB-LB^n}{UB} < \epsilon$ where $\epsilon > 0$ is a pre-defined tolerance level, terminate the algorithm and the best solution found is $(\bar{X}, \bar{Y}, \bar{S})$. Otherwise, continue to Step 3.

Step 3: Update the dual multipliers using the subgradient procedure using information from the lower bound solution $(\underline{X}^n, \underline{Y}^n, \underline{S}^n)$. Let the new multipliers be $(\theta^{n+1}, \pi^{n+1})$ and increment n by 1. If n is larger than a predefined iteration limit, terminate. Otherwise go to Step 1.

5. COMPUTATIONAL RESULTS

5.1. Experimental Setup

In this section, we describe the computational experiments conducted to test the performance of our solution procedure. Three data sets from Daskin [10] with 49, 88 and 150 US cities (with 1990 Census data) respectively are used. The cost coefficients are adjusted in the following manner: The fixed costs f_j are unchanged from the Daskin [10] dataset. The transportation costs (d_{ij}) are obtained by dividing the distance between two nodes from Daskin [10] by 10. Demand rates (λ_i) are set to 10^{-6} times the population figure given in Daskin [10]. The transportation lead time (α_j) is set to be equal to 1/10 of the distance between the corresponding demand node and Springfield, IL, in the 49-node dataset and Chicago in the 88-node and 150-node datasets. Finally, the per-unit holding and penalty costs (h_0, h_j, p) are set to 50 and 150, respectively. The algorithm is coded in C++ and run on a PC with Intel Core 2 Duo CPU (2.13 GHz), 2 GB of RAM, and Windows Vista.

5.2. Algorithmic Performance

In the first experiment, we are concerned with the performance of the algorithm, i.e., speed and optimality gap. We

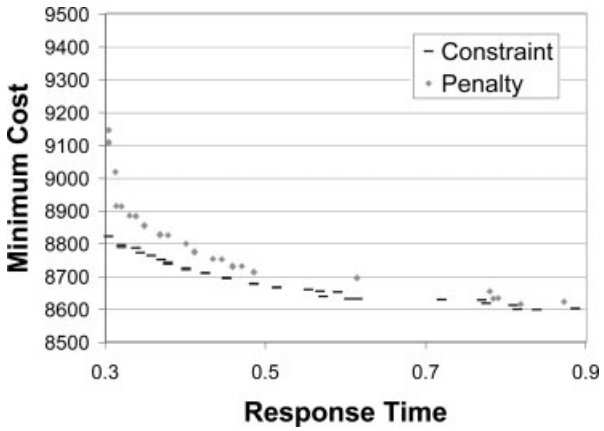

Figure 1. Optimal costs of using penalty cost and service constraint to achieve a certain average response time.

run the algorithm with the three data sets for different values of plant and SC capacities (C_0, $C_j = 10$ or 5), utilization rate ($\rho = 0.9$ or 0.5), response time requirement ($\tau = 5.5$ or 1.5), and distance requirement ($d_{\max} = 2000$ or 500 miles).

For all instances tested, the algorithm converges to a feasible solution within approximately 1 to 2% of optimality in reasonable amount of time. In the majority of the instances, the optimality gap is less than 1%. For the 49-node data set, the solution time range from a few seconds up to approximately 3 min. All 88-node instances are solved within 2 min. The solution times for the 150-node instances range from a minute up to 35 min. The solution times recorded are reasonable for the respective problem sizes.

5.3. Effects of Response Time Requirement and Penalty Cost

In the second experiment, we wish to test the impact of the response time requirement and the backorder penalty cost and their relationship. In traditional inventory problems with service considerations, it can be shown that there is a one-to-one correspondence between the service level and the penalty cost. In reality, it is often difficult to quantify "penalty costs" of having a stock-out or not fulfilling an order within a specified time. We would like to see whether a response time constraint can act as a perfect substitute for a penalty cost in our problem.

Figure 1 shows how high the total cost of location, transportation, and inventory holding needs to be (vertical axis) to maintain a expected response time at a certain level (horizontal axis), by using a penalty cost or a tight response time requirement. The figure shows that the curve obtained by using penalty costs lies above the one obtained by using the response time requirement. This means that setting a tight response time requirement can ensure a certain response time in a more cost-effective manner than by using penalty costs.

In our solution algorithm, the response time constraints (5) are relaxed by adding penalty terms to the objective function. We notice that this is structurally equivalent to increasing the penalty cost p by the dual multiplier θ_j. As the Lagrangian relaxation algorithm converges, the optimal dual multiplier values would almost ensure that the constraints are satisfied. Therefore, we would expect the following two instances: the first one with 0 penalty cost and a tight response time constraint, and the second one with very loose response time constraint and unit penalty cost equal to the optimal θ_j values found in solving the first one, to give optimal solutions that are almost the same. This means that penalty costs and response time requirement are to some degree, equivalent in our model. However, this is not true in the practical sense. We do not know the optimal dual multipliers for the instance with tight response time constraints before solving it. Even if the values are known, they are generally not equal across candidate SC locations. However, there is no reason to define different unit penalty costs for different candidate SC locations, before even deciding which SCs will be opened and which customers they will serve. Our computational experiment shows that using a uniform (across candidate SC locations) penalty cost parameter will require higher costs to achieve a certain response time level than using a tight response time requirement.

5.4. Benefits of Using Integrated Model

Following the traditional approach adopted by logistics planners, it is natural to solve the location and inventory problems sequentially. In this section, we compare the solutions generated by our integrated approach with those obtained from the traditional sequential approach. To do so, we generate a number of instances by varying the unit holding and backorder costs and generate two solutions. The integrated solution is obtained from solving our model, and the sequential solution can be obtained by first solving an uncapacitated facility location (UFL) problem (ignoring inventory effects) and then optimizing the inventory levels given the facility locations.

Figure 2 shows comparisons of the two solutions. We first observe that as expected, the integrated approach gives a lower total cost. Similar results are reported by Shen and Qi [31] and Candas and Kutanoglu [7]. For a closer look at the differences, we plot the base stock levels of the two solutions at both echelons. It can be seen that the sequential approach always places more inventory at the plant but less at the SCs compared to the sequential solution. One possible explanation is that because the uncapacitated facility location problem only considers the transportation cost, the sequential solution tends to locate more SCs relatively closer to

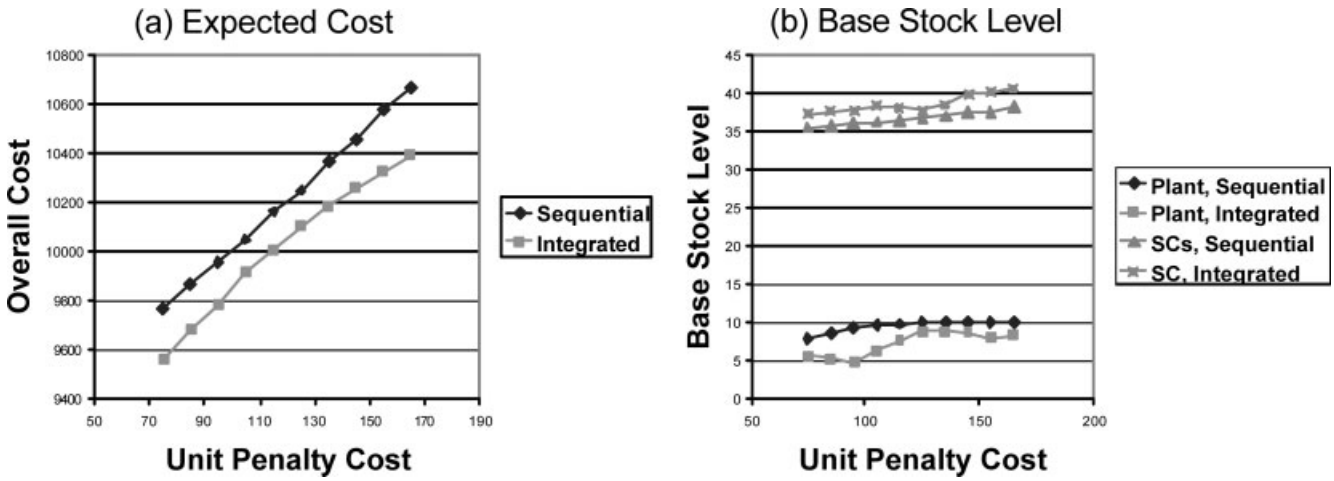

Figure 2. Optimal costs and base stock levels of using integrated and sequential approaches.

customers, each of which handling a smaller demand. Therefore, the degree of inventory risk-pooling at the SCs is small. The result is that a larger proportion of inventory is transferred to the plant (which serves all demand) such that risk-pooling is achieved. This is less responsive to customer demand than the integrated solution because less inventory is stored at SCs that serve customers directly.

5.5. Comparisons with One-Echelon Systems

It is often convenient to manage inventory at a single level. Benjaafar et al. [2] consider the service parts inventory-location problem where inventory can only be held at the lower echelon (the SCs). In this experiment, we would like to study the benefits from the possibility of holding inventory at the plant, also. Note that our formulation models a single-echelon system when the plant storage capacity C_0 is set to 0.

In the experiment, we allow the response time requirement τ to vary between 0.3 and 1.3. For each level of τ, we solve the problem with plant storage capacity C_0 equal 0 (one-echelon) and 10 (two-echelon). The storage capacity at the SCs, C_j, is set to 10. We plot the expected cost against the response time requirement in Figure 3.

From Figure 3, we observe that when the response time requirement is loose (high value of τ), it is optimal to operate a single-echelon system and not to stock at the plant. The two-echelon solutions have plant base stock levels equal to 0. However, when the response time requirement becomes tight, it is beneficial to keep inventory at the plant and manage a two-echelon system. Keeping inventory at the plant allows fewer units to be kept at the SCs. Although the total inventory in the system may increase, the increased inventory at the plant is shared among all SCs. This sharing enhances availability and therefore the same response

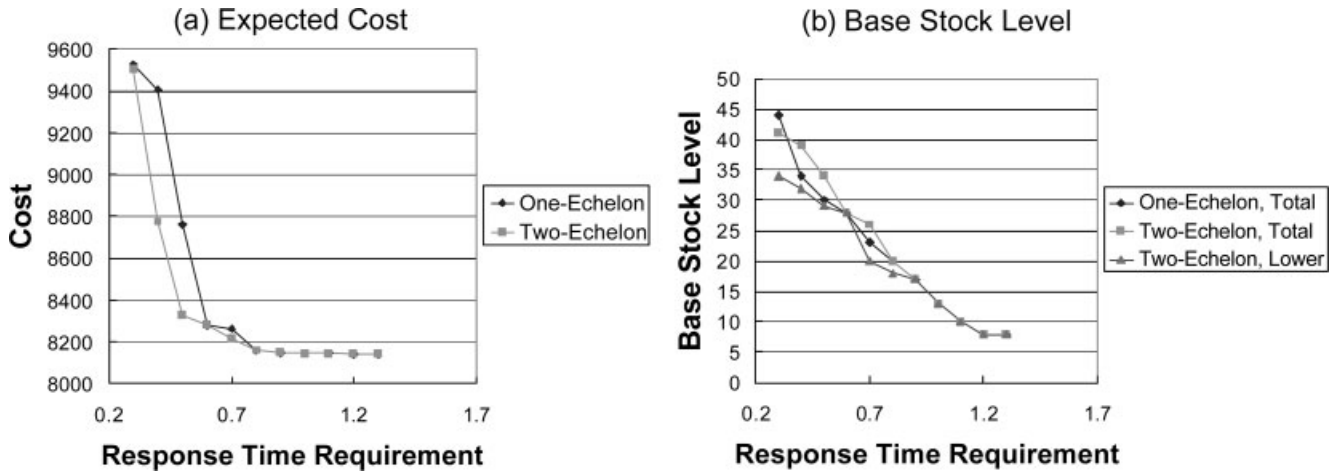

Figure 3. Costs of one-echelon and two-echelon systems.

time requirements can be met with lower costs. For some instances (e.g., when $\tau = 0.4$), the expected cost can increase by 7% if keeping stock at the upper echelon is not allowed.

5.6. Scenario-Based Model

Facility location involves long-term decisions. One challenge that often arises in strategic supply chain design is the unavailability of accurate demand and cost forecasts at the time of making design decisions such as facility location. As such decisions are difficult and very costly to reverse, it is important that they are robust against the uncertain environment. Snyder [37] surveys various location models that aim at producing robust or reliable solutions. Scenario-based modeling is a very popular technique for this class of problems. Under this approach, uncertainty is characterized by a finite set of discrete scenarios. One example is the article by Snyder et al. [38], who extend the inventory-location model of Shen et al. [29] to incorporate uncertainty in the business environment (i.e., demand and shipping costs).

In this section, we assume that the demand rates (λ_i) are uncertain and formulate a two-stage stochastic programming model to generate robust solutions. It is straightforward to extend the model and allow other parameters such as shipping, holding and penalty costs to be uncertain as well. The first-stage decision is to determine the subset of candidate sites at which to locate SCs (i.e., the X_j variables) given a finite set of possible scenarios of demand rates. Then in the second stage, the company decides the customer-SC allocations and base stock levels after observing which one of the scenarios is realized.

Define K as the set of scenarios and q_k as the probability that scenario $k \in K$ is realized. Moreover, we add a subscript k to any parameter (e.g., λ_{ik}) and decision variable (e.g., S_{jk} that may take on different values under different scenarios. Then we may formulate the scenario-based problem as follows:

$$\begin{aligned}
\min & \sum_{j\in J} f_j X_j \\
& + \sum_{k\in K} q_k \left\{ \sum_{j\in J} \left[+ \sum_{i\in I} \left[p\left(\frac{\rho^{S_{0k}+1}}{\lambda_k(1-\rho_k)} + \alpha_j \right) + d_{ij} \right] \lambda_{ik} Y_{ijk} \right.\right. \\
& \qquad \left. + (h_j + p) \sum_{s=0}^{S_{jk}-1} F_{jk}(s) - p S_{jk} \right] \\
& \qquad \left. + h_0 \left[S_{0k} - \frac{\rho_k}{1-\rho_k}\left(1 - \rho_k^{S_{0k}}\right) \right] \right\}
\end{aligned}$$

Subject to:

$$\begin{aligned}
\sum_{j\in J} Y_{ijk} &= 1, \quad \text{for each } i \in I, k \in K \\
Y_{ijk} &\le X_j, \quad \text{for each } i \in I, j \in J, k \in K \\
S_{jk} &\le C_j X_j, \quad \text{for each } j \in \{0\} \cup J, k \in K \\
X_j &\in \{0,1\}, \quad \text{for each } j \in J \\
Y_{ijk} &\in \{0,1\}, \quad \text{for each } i \in I, j \in J, k \in K \\
S_{jk} &\ge 0, \text{ integer}, \quad \text{for each } j \in \{0\} \cup J, k \in K
\end{aligned}$$

We apply a multiobjective technique and consider two objectives: the expected cost and the worst-case "service time" across all scenarios. The "service time" for a customer under a scenario considered in this section is the sum of the expected response time at the assigned SC plus the deterministic transportation lead time from the SC to the customer. We attempt to minimize the longest service time among all customers across all scenarios. By considering the worst-case service time over all possible scenarios, this objective reflects the desire for the supply chain to be robust. Note that the service time objective replaces the service time and distance constraints (21 and 23) that are included in the single-scenario model.

We apply a genetic algorithm similar to the one described in Shen and Daskin [30]. For the 49-node data set, 30 scenarios are generated by perturbing the demand rates. Specifically, the demand rate for customer i under scenario k is given by $\lambda_{ik} = \lambda_i(0.5 + \epsilon_{ik})$, where λ_i is the demand rate for customer i used in the previous tests and ϵ_{ik} is a $[0, 1]$ random variable drawn independently for each (i, k) pair. We assign equal probability to all scenarios.

Figure 4 shows the efficient frontier (i.e., those that are not worse than any other solution in *both* objectives) after 25, 50, 100, and 500 solutions. It can be observed that the efficient frontier moves toward the bottom-left corner (i.e., improves) as the number of generations increases. Similar to the results in Section 6.4, we observe that the slope of the efficient frontier is steeper when the expected cost is low. This suggests that starting from a solution with low average-case cost, the marginal cost of improving the worst-case service time is small. For instance, improving the worst-case service time by a factor of 3 (1.8 to 0.6) only increases the average-case cost by about 20% (8000 to 10,000).

Furthermore, we observe that the improvement of the efficient frontier appears to diminish with the number of generations. The relevant region (the bottom-left region) of the 100-generation frontier approximates the 500-generation frontier closely. Therefore, it may be sufficient to run a small number of generations to obtain fairly good results. The running time for 100 generations is less than 10 min.

Before ending this section, we note that the model discussed in this section can be applied to multi-commodity

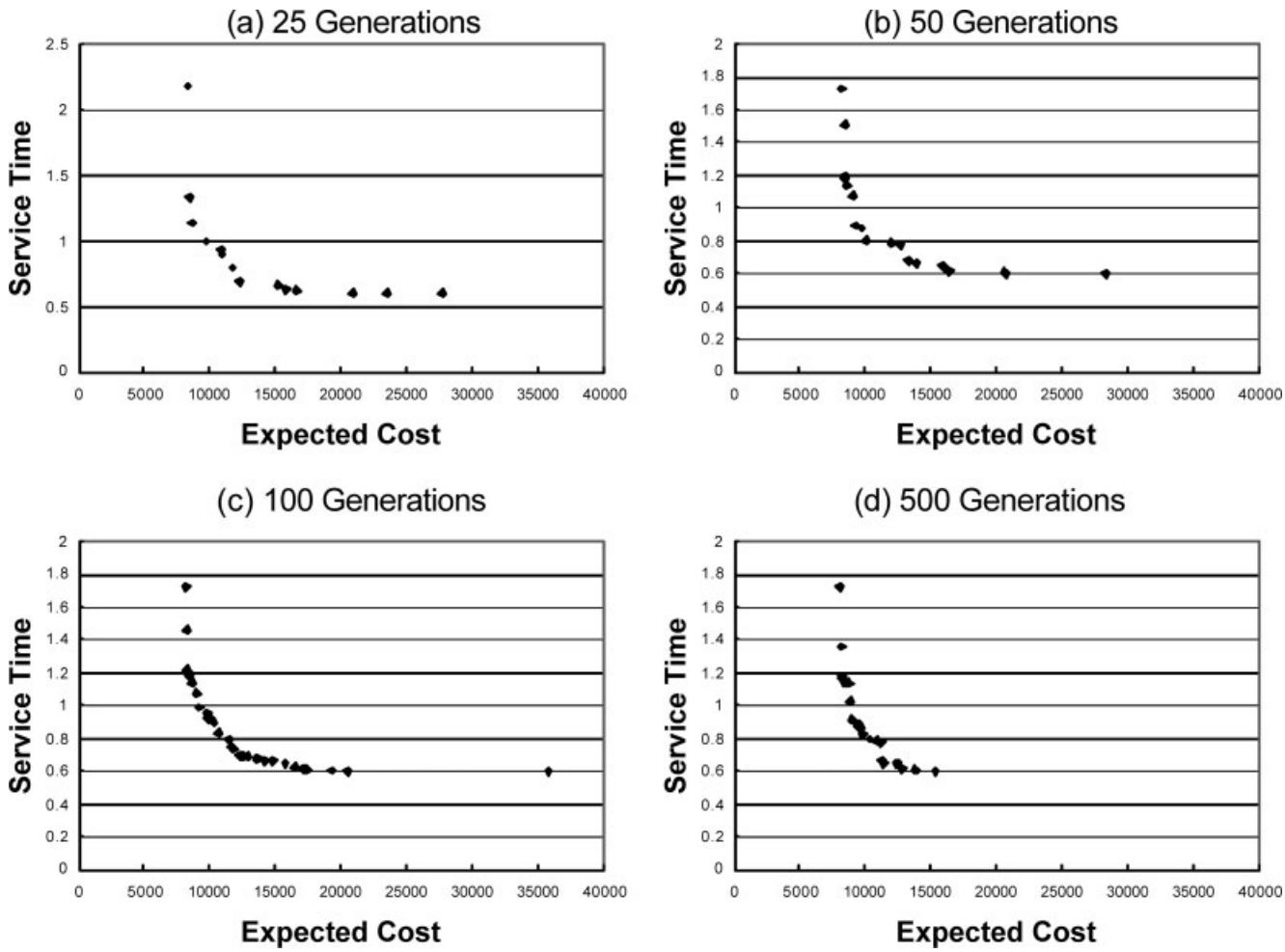

Figure 4. Trade-off curve of expected cost vs. service time for scenario-based problem.

problems. Service parts systems often handle multiple non-substitutable parts. By interpreting K as the set of commodities, the demand rates (λ_{ik}) as well as the assignment and inventory variables (Y_{ijk}, S_{jk}) become commodity-specific. Then the same formulation is equivalent to locating common SCs to store and distribute multiple commodities. For each commodity, the customer-SC assignment and the base stock levels at the plant and SCs are optimized. The same GA can be applied to produce high-quality solutions.

6. CONCLUSIONS AND FUTURE RESEARCH

In this article, we formulate a model for designing of a service parts network to ensure short response times. Fixed costs of locating service centers, shipping costs, and inventory holding costs at both the plant and the service centers are considered. The model is formulated as a nonlinear integer programming problem. The formulation can be viewed as an extension of the uncapacitated fixed-charge location problem.

The inventory-related costs and service performance metrics are approximated using ideas similar to those in Sherbrooke [32]. As a result, we are able to decompose the problem into a series of convex optimization problems using Lagrangian relaxation. Then the algorithm is tested using datasets with 49, 88, and 150 nodes. Computation times are relatively short for such a complex integer nonlinear programming problem, reflecting the efficiency of the proposed solution approach. In addition, we compare the difference between using a response time requirement and penalty costs in our model. We demonstrate that by using a response time requirement, it is less costly to achieve a target response time than by using penalty costs as a surrogate. Finally, we highlight the importance of selecting the right trade-off between service and cost. The marginal cost for improving response time is typically higher when the response time requirement is tight than when it is loose. Therefore, care should be taken in

specifying response time requirements to avoid unnecessary costs. Our computational results also demonstrate the cost savings from implementing a two-echelon system as opposed to a single-echelon one when response time requirements are tight. By storing some inventory at the plant that is shared by all SCs, it is possible to achieve the same response time requirement at a lower cost.

Finally, we formulate a scenario-based model and propose a genetic algorithm to solve it. Using two contrasting objectives of expected cost and worst-case expected response time among the SCs, our model captures the trade-off between average-case performance and robustness in view of future uncertainty. The same model and solution approach can also be applied to solving multiple-commodity problems which are common in service parts systems.

We plan this work in several directions. First, we would like to relax the assumption of the $(S-1, S)$ inventory policy to allow batch ordering by SCs. This extension will make the model applicable to a larger variety of supply chain design environments. Second, we would like to consider the case where parts of the network may be disrupted. In many supply chain and military settings, disruptions may occur and certain facilities (the plant or service centers) or arcs may break down. We would like to address important questions such as how to design robust systems that are able to operate under unanticipated breakdowns of nodes or arcs on the network.

Another important issue is the nature of the response time requirement. The model we present in this article considers only the mean response time. However, it is desired that service be not just quick on average but also reliable. In classical inventory models, service levels are usually defined using probabilistic constraints (e.g., type I service levels in continuous review problems). As an extension, we would like to redefine the service requirements using either chance constraints on the response time distribution or constraints on the service time variance.

ACKNOWLEDGMENTS

The authors would like to thank Professor Candace Yano, Professor Mark Daskin, the associate editor, and the anonymous referees for their helpful comments and suggestions. This research was supported in part by NSF CAREER award DMI-0621433.

REFERENCES

[1] S. Axsäter, Inventory control, second edition, Springer, New York, 2006.
[2] S. Benjaafar, Y. Li, D. Xu, and S. Elhedhli, Demand allocation in systems with multiple inventory locations and multiple demand sources, Manufacturing Service Oper Manage 10 (2008), 43–60.
[3] R.P. Brent, Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[4] J.A. Buzacott and J.G. Shanthikumar, Stochastic models of manufacturing systems, Prentice Hall, New Jersey, 1993.
[5] D. Caglar, A multi-echelon spare parts inventory system with emergency lateral shipments subject to a response time constraint, Ph.D. Thesis, Department of Industrial Engineering and Management Sciences, Northwestern University, IL, 2001.
[6] D. Caglar, C.L. Li, and D. Simchi-Levi, Two-echelon spare parts inventory system subject to a service constraint, IIE Trans 36 (2004), 655–666.
[7] M.F. Candas and E. Kutanoglu, Benefits of considering inventory in service parts logistics network design problems with time-based service constraints, IIE Trans 39 (2007), 159–176.
[8] M.A. Cohen, P.A. Kleindorfer, and H.L. Lee, Near-Optimal Service Constrained Stocking Policies for Spare Parts, Oper Res 37 (1989), 104–117.
[9] M.A. Cohen, Y.S. Zheng, and V. Agrawal, Service Parts Logistics: A Benchmark Analysis, IIE Trans 29 (1997), 627–639.
[10] M. Daskin, Network and Discrete Location: Models, Algorithms, and Applications, John Wiley Sons, New York, 1995.
[11] M. Daskin, C. Coullard, and Z.J. Shen, An Inventory-Location Model: Formulation, Solution Algorithm and Computational Results, Annals Oper Res 110 (2002), 83–106.
[12] Z. Drezner (Editor), Facility Location: A Survey of Applications and Methods, Springer, New York, 1995.
[13] Z. Drezner and H.W. Hamacher (Editors), Facility Location: Applications and Theory, Springer, New York, 2002.
[14] E. Eskigun, R. Uzsoy, P.V. Preckel, G. Beaujon, S. Krishnan, and J.D. Tew, Outbound Supply Chain Network Design with Mode Selection, Lead Times and Capacitated Vehicle Distribution Centers, Eur J Oper Res 165 (2005), 182–206.
[15] E. Eskigun, R. Uzsoy, P.V. Preckel, G. Beaujon, S. Krishnan, and J.D. Tew, Outbound Supply Chain Network Design with Mode Selection and Lead Time Considerations, Naval Res Log 54 (2007), 282–300.
[16] M. Ettl, G.E. Feigin, G.Y. Lin, and D.D. Yao, A Supply Network Model with Base-Stock Control and Service Requirements, Oper Res 48 (2000), 216–232.
[17] S.C. Graves, A Multi-Echelon Inventory Model for a Repairable Item with One-for-one Replenishment, Manage Sci 40 (1985), 597–602.
[18] IBM (March 16, 2009), Spare Parts Inventory Management Solution—Business View. Available at: http://www-03.ibm.com/industries/automotive/us/detail/solution/U225671Y82885J57.html?tab=2, Accessed date: March 16, 2009.
[19] R.R.P. Jackson and D.G. Nickols, Some Equilibrium Results for the Queuing Process E_k/M/1, J R Stat Soc Ser B 18 (1956), 275–279.
[20] E. Kutanoglu, Insights into Inventory Sharing in Service Parts Logistics Systems with Time-based Service Levels, Computers Ind Eng 54 (2008), 341–358.
[21] H.L. Lee and C. Billington, Material Management in Decentralized Supply Chains, Oper Res 41 (1993), 835–847.
[22] MacroSys Research and Technology (Aug 25, 2005), Logistics Costs and U.S. Gross Domestic Product, Available at: http://ops.fhwa.dot.gov/freight/freight_analysis/econ_methods/lcdp_rep/index.htm, accessed date: April 5, 2007.
[23] K. Moinzadeh and H.L. Lee, Batch Size and Stocking Levels in Multi-Echelon Reparable Systems, Manage Sci 32 (1986), 1567–1581.

[24] L.K. Nozick and M.A. Turnquist, A Two-Echelon Inventory Allocation and Distribution Center Location Analysis, Transportation Res E 37 (2001), 425–441.
[25] L. Ozsen, M. Daskin, and C. Coullard, Capacitated Facility Location Model with Risk Pooling, Naval Res Log 55 (2008), 295–312.
[26] L. Ozsen, M. Daskin, and C. Coullard, Facility Location Modeling and Inventory Management with Multi-Sourcing, Transportation Sci (in press) (2008).
[27] Z.J. Shen, A Multi-Commodity Supply Chain Design Problem, IIE Trans 37 (2005), 753–762.
[28] Z.J. Shen, Integrated Supply Chain Design Models: A Survey and Future Research Directions, J Industrial Manage Optim 3 (2006), 1–27.
[29] Z.J. Shen, C. Coullard, and M.S. Daskin, A Joint Location-Inventory Model, Transportation Sci 37 (2003), 40–55.
[30] Z.J. Shen and M.S. Daskin, Trade-offs Between Customer Service and Cost in Integrated Supply Chain Design, Manufacturing Service Oper Manage 7 (2005), 188–207.
[31] Z.J. Shen and L. Qi, Incorporating Inventory and Routing Costs in Strategic Location Models, Eur J Oper Res 179 (2006), 372–389.
[32] C.C. Sherbrooke, METRIC: A Multi-Echelon Technique for Recoverable Item Control, Oper Res 34 (1968), 311–319.
[33] J. Shu, C.P. Teo, and Z.J. Shen, Stochastic Transportation-Inventory Network Design Problem, Oper Res 53 (2005), 48–60.
[34] D. Simchi-Levi and Y. Zhao, Safety Stock Positioning in Supply Chains with Stochastic Lead Times, Manufacturing Service Oper Manage 7 (2005), 295–318.
[35] C.C. Sherbrooke, Optimal Inventory Modeling of Systems, Kluwer Academic Publishers, Norwell, MA, 2004.
[36] K. Sourirajan, L. Ozsen, and R. Uzsoy, A Single-Product Network Design Mode with Lead Time and Safety Stock Considerations, IIE Trans 39 (2007), 411–424.
[37] L. Snyder, Facility Location Under Uncertainty: A Review, IIE Trans 38 (2006), 537–554.
[38] L. Snyder and M.S. Daskin, Reliability Models for Facility Location: The Expected Failure Cost Case, Transportation Sci 39 (2005), 400–416.