Benefits of Considering Inventory in Service Parts Logistics Network Design Problems with Time-based Service Constraints

Mehmet Ferhat Candas and Erhan Kutanoglu *
Operations Research and Industrial Engineering Graduate Program
Department of Mechanical Engineering
The University of Texas at Austin

February 6, 2006

Abstract

We study the integrated logistics network design and inventory stocking problem as characterized by the interdependency of the design and stocking decisions in service parts logistics. These two sets of decisions have been usually considered sequentially in practice, and the associated problems have been tackled separately in the research literature. The overall problem is typically further complicated due to time-based service constraints that provide lower limits for the percentages of demand satisfied within specified time windows. We introduce an optimization model that explicitly captures the interdependency between network design (location of facilities, and allocation of demands to facilities) and inventory stocking decisions (stock levels and their corresponding stochastic fill rates), and present computational results from our extensive experiments that investigate the effects of several factors including demand levels, time-based service levels, and costs. We show that the integrated approach can provide significant cost savings over the decoupled approach (solving the network design first and inventory stocking next), shifting the whole efficient frontier curve between cost and service level to superior regions. We also show that the decoupled and integrated approaches may generate totally different solutions, even in the number of located facilities and in their locations, magnifying the importance of considering inventory as part of the network design models.

*Corresponding author: erhank@mail.texas.edu
1 Introduction

Increasing worldwide competition and shrinking profit margins have been forcing high technology product manufacturers to differentiate themselves from others in different ways. Providing fast, high-quality after-market service is an important way to achieve this. After-market service is providing necessary service and replacement parts to existing, geographically dispersed customers when they experience any problem with their product. The service is provided as part of the contract between the customer and the manufacturer, therefore designing and operating a logistics network capable of serving customers in a time-responsive manner is crucial for after market service in which service parts logistics is a critical part. Service parts logistics (SPL) includes activities such as designing a responsive network of part stocking facilities, deciding inventory ordering policies, stocking parts, and dispatching the required parts from facilities to the customers in need. A major challenge in SPL is to provide the service and satisfy customer's request within the committed time. Latest surveys (Cohen et al., 1997; Poole, 2003) show that 40 to 50% of profits of manufacturers come from parts, maintenance, and servicing, which makes SPL to be a $21 billion industry.

Locating inventory stocking facilities, allocating customer demands to these facilities and selecting stock levels maintained at these facilities are main decisions while designing the SPL system. Traditionally, location and allocation decisions (collectively called logistic network design or LND) are considered part of strategic and long-term decisions, which are typically made before any tactical decisions such as inventory levels. However, redesigning an existing network more frequently is becoming more likely due to outsourced warehousing and delivery services common in SPL. For example, companies such as UPS and FedEx provide access to extensive global logistics networks and offer services ranging from sourcing of parts from vendors to shipments to customer sites. This allows companies using third party logistics services to expand or shrink their network as needed without much difficulty. Moreover, due to the time-based service level requirements that are critical part of any SPL system, there is a stronger interaction between “strategic network design” and “tactical” inventory decisions as the service requirements are not only a coverage issue (whether a customer's demand is covered by a nearby facility), but also a function of the part availability at that facility. Thus, we conjecture that considering the effects of network decisions on inventory (and vice versa) in an integrated model becomes critical for optimization of an SPL system.

We note that the ultimate decision made in the introduced model is still designing the network. We make inventory stocking part of the network design model to make better overall decisions and find a network design that will not only minimize facility and transportation costs but also minimize inventory costs while ultimately capturing service level tradeoffs. The inventory decisions can and should be updated over time as needed without overhauling the network. Motivated by these real challenges in today's SPL systems, we model the integrated network design and inventory stocking
problem. We explicitly consider inventory decisions and costs in the LND problem, which is itself already complicated due to explicit time-based service constraints. We also quantify the benefit of considering both network design and inventory decisions in the same model and identify the conditions and problem settings where this benefit is significant.

2 Literature Review

We review papers from (1) the location/allocation/network design literature, (2) the multi-location service constrained inventory management, and, of course, (3) a limited number of papers investigating similar integration issues.

2.1 Facility location/network design problems

Facility location and network design problems with countless variations have been studied extensively in the literature. The relevant papers in this area study service constrained, stochastic, or reliability-based problems. There are studies that address global uncertainties such as exchange rates, transfer prices, taxes and market prices along with supplier reliability and lead time uncertainty in a single-echelon model (Vidal and Goetschalckx, 2000). These studies define reliability as the probability of being on time by all suppliers and the amount of product shipped is at least a specified target value. Another related area is the study of fixed-charge location/allocation problems with demand/service coverage restrictions. Typical examples here include emergency facility location problems (e.g., Goldberg and Paz (1991)) and similar logistics applications (e.g., Nozick (2001)). Snyder and Daskin (2005) investigate the uncapacitated facility location (UFL) problem with potentially unreliable facilities, where a customer may not be served from its facility due to “failure” of the facility which occurs with certain probability. Two recent reviews (Daskin et al., 2005; Snyder, 2004) in this area surveys more papers in this area in the supply chain management context. Daskin (1995) is a reference text on discrete facility location problems. Magnanti and Wong (1984) review the early literature on the facility location problems, while Drezner (1995) summarizes the overall research effort by 1995.

2.2 Inventory Management and Service Parts Logistics

Similar to the location literature, there is a vast amount of work on inventory management. We only review the ones that are most relevant to our work, and refer to Zipkin (2000) for a recent text on the topic. A stream of research related to the inventory-portion of our study explicitly considers service level constraints. We can list Chen and Krass (2001) for a single facility operating reorder point and order-up-to level policy \((s, S)\), and Agrawal and Seshadri (2000) for order-quantity and reorder point policy \((Q, r)\). One example (Song, 1998) investigates a simplified time-based service level for base stock policies for a group of items. For an example of a multi-facility inventory
allocation problem, see (Rappold and Muckstadt, 2000).

Early literature on spare/service parts inventory management in multi-echelon systems includes (Sherbrooke, 1968, 1986), and (Muckstadt, 1973). Successful applications in service parts logistics systems include such industries as automotive (Cohen et al., 2000), computer and other electronic equipment service (Cohen et al., 1988, 1990, 1999), and military (Rustenburg et al., 2001). One of the early works on multi-echelon service parts inventory management is by Muckstadt and Thomas (1980). A limited number of studies in this group consider fill rate-based service allocation (Cohen et al., 1988, 1989, 1992). While the book by Sherbrooke (1992) provides an overall review of multi-echelon inventory management from a military perspective with a focus on repairable parts, a recent text by Muckstadt (2005) is now a reference for general SPL research.

2.3 Network Design Models with Inventory Considerations

Integration of facility location and inventory problems is a very recent research area, hence there are only a handful of papers on this topic. As they are most relevant to our study, we review them in more detail. One of the early papers modifies the uncapacitated facility location problem to implicitly consider limited inventory levels (Barahona and Jensen, 1998). Nozick and Turnquist (1998) approximate inventory costs as part of the fixed facility costs assuming a linear relationship between inventory and the number of open facilities, and propose a model that takes service coverage as a constraint. Nozick and Turnquist (2001) extend this model and treat demand coverage as part of the objective function. The paper by Daskin et al. (2002) is probably the first study that explicitly includes inventory costs as part of a simple, uncapacitated facility location model. Their model assumes economic order quantity (EOQ)-based ordering and constant fill rate-based safety stocks across all facilities. The total cost function including the inventory related terms makes the overall model a nonlinear integer program which is then solved using Lagrangian relaxation. (A parallel and a very similar model is analyzed in (Miranda and Garrido, 2004).) Shen et al. (2003) develop a column generation-based method to solve the same model, while Ozsen et al. (2004 (Submitted to Networks) extend the model to include facility capacities. The other extensions of the same model include a scenario-based stochastic version (Snyder et al., 2003), an approximation algorithm for the version that ignores transportation costs (Teo et al., 2001), and a version with customer-specific service levels (Shen and Daskin, 2005). Another relevant paper is a grid-based location-inventory model (Erlebacher and Meller, 2002), In contrast to these important contributions, our model tries to achieve a system-level service by allocating it to multiple facilities in an optimal manner, hence considering the varying fill rates as explicit decision variables. Moreover, our model is especially applicable to low-demand settings with one at a time ordering, whereas the papers cited here model inventory as an objective-changing cost component with EOQ-ordering, making them more applicable to high-demand settings with significant order setup costs.
3 Problem Definition and Modeling

We first define the problem setting and introduce the notation, and then introduce the integrated model formulation, dealing with its complexities along the way.

3.1 Problem Setting and Notation

For a given set of customers and their mean demands, we seek to (1) locate a set of stocking facilities selected from a set of candidate facility locations, (2) allocate customer demands to these located (open) facilities, and (3) determine stock levels to be used at the open facilities. Decision sets (1) and (2) make up the network design, and (3) are the inventory stocking decisions. The goal is to make these decisions with minimum possible total facility, transportation, and inventory costs while achieving the target (required) service levels. We make the following assumptions that facilitate the model development:

- We assume that network design involves the stocking facilities that are all in one echelon facing the direct demand from geographically dispersed customers. We assume that these facilities to be located are replenished from a central warehouse with infinite capacity (that is, the central warehouse can replenish the stocking facilities anytime without any delay). The lead times from the central warehouse to all facilities is the same, known and constant.
- Due to the low-demand nature of the motivating SPL problem, we assume that the facilities use continuous review, one-for-one (or base-stock, also called \((S - 1, S)\)) replenishment policy. This is typical as demands are low, and lead times are relatively short in SPL systems. For examples, all the models in (Sherbrooke, 1992) use this policy, even for higher-echelon facilities, where demands from lower-echelon facilities are aggregated.
- We assume that demand for each part at each demand point arrives one at a time according to an independent Poisson process, typical in low-demand settings. This is a common assumption in the SPL literature (see, e.g., Sherbrooke (1992); Muckstadt (2005)). We assume that we know the mean demand rates obtained from the part failure rate distributions and the number of parts used at each demand point. Any unsatisfied demand due to a stockout at a facility is backordered.
- We assume that service contracts, demand aggregation across individual customers to form demand points, and service level aggregation to obtain target service level for the region are done a priori. This translates into a percentage of demand to be satisfied within a certain service time window. For example, a typical aggregate service level may read “70% of total demand for part 1 must be satisfied from facilities that are within 4 hours of the demand points.”
- For simplicity in presentation, we assume just one service time window. Extending the model (as will be seen) for multiple windows is straightforward. In the experiments, however, we vary the time window as a control factor to see its effect on the results.
We assume that we know which customers a facility can serve within the service time window. As this is usually a function of distance and the mode of transportation available to the facility and customer, we assume that this processing of transportation times is performed for each customer and facility pair a priori. We further assume that each customer’s part request is satisfied by a single direct shipment from a facility, without any shipment consolidation or bundling. Not only is this actual practice in SPL systems, but also it is very unlikely to have time or opportunity to consolidate multiple shipments due to low demands and strict time windows.

We finally assume that a part’s demand can be satisfied by more than one facility. Due to the structure of the problem, this does not happen very often in the optimal solutions, but it provides opportunities to find superior inventory stocking level combinations. At the operational level, this means assigning the demand to the facilities randomly with probabilities corresponding to their long-run fractions found in the optimal solution. We assume that the facility satisfies the demands in first come first served order regardless of where the requesting demand point is. With this, total demand assigned to each facility follows a Poisson distribution with the corresponding aggregate mean demand. We assume that there is no inventory transfer or transshipments between the stocking facilities. Although some SPL systems use transshipments in case of stockouts to improve service and pool risk, calculating stock levels even for a given fixed facility locations with known demands is a nontrivial task Cohen et al. (1986); Alfredsson and Verrijdt (1999). As we show here, the new integrated network design and inventory model without transshipments is rich enough to lead to interesting results. In that sense, this paper serves as a precursor to study more complex models with transshipments.

We now introduce the notation. We are given a set of candidate facility locations $I$ (indexed by $i$), a set of demand points $J$ (indexed by $j$), and a set of parts $K$ (indexed by $k$). When we open facility $i$ (or more correctly, locate a facility at candidate location $i$), we incur an annual cost of $f_i$ for operating the facility. The unit transportation cost between facility $i$ and demand point $j$ for part $k$ is $c_{ijk}$. Let $\tau_{ij}$ be the transportation time from facility $i$ to demand point $j$. Comparing $\tau_{ij}$ to the service time window $w$, we obtain $\delta_{ij}$, the identifier which takes value 1 if facility $i$ can ship a part requested at demand point $j$ within the specified service time window ($\tau_{ij} \leq w$), 0 otherwise ($\tau_{ij} > w$). The mean annual demand rate is $d_{jk}$ for part $k$ at demand point $j$ (i.e., the rate at which demand point $j$ experiences part $k$ failures). With a little notation abuse, we denote the total annual mean demand rate for part $k$ (across all customer demand points in the region) by $d_k$, $d_k = \sum_{j \in J} d_{jk}$. The target level for the system-wide service for the specified service time window is $\alpha_k$ for part $k$. Hence, 100$\alpha_k$% of the total annual demand for part $k$ (of all part $k$ failures) needs to be satisfied within the time window.

We introduce the development of the integrated model that simultaneously makes network design and inventory stocking decisions, and captures the true relationship among location, demand al-
location, fill rate-based part availability, and system-wide service levels. Using a classical facility location formulation as a base, we first build an intuitive, but highly nonlinear integer programming model (Section 3.2). We then show how one can take advantage of the low-demand rates (hence, low stock levels) to linearize the model (Section 3.3). We finally add a post-processing stage to eliminate the drawbacks of the linearization and to refine the model’s solution (Section 3.4).

3.2 Nonlinear Integer Programming Model

We start by developing a traditional network design model based on modifications of the (fixed-charge, multi-commodity) uncapacitated facility location problem, the main change being the addition of the service coverage constraints. The model uses the following decision variables:

- \( X_{ijk} \) = long run fraction of part \( k \)’s demand at demand point \( j \) allocated to facility \( i \)
- \( Y_i = 1 \) if facility \( i \) is open, 0 otherwise

With the allocations \( X_{ijk} \), we define the total demand rate for facility \( i \) for part \( k \) as follows:

\[
d_{ik} = \sum_{j \in J} d_{jk} X_{ijk}. \tag{1}
\]

The demand rate at which facility \( i \) experiences part \( k \) failures from customers within the time window is

\[
d_{ik}^{\leq w} = \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk}. \tag{2}
\]

Hence, the fraction of part \( k \) failures from customers within the time window for facility \( i \) is \( d_{ik}^{\leq w} / d_{ik} \). Moreover, the fraction of part \( k \) failures that facility \( i \) addresses is \( d_{ik} / d_k \) out of all regional demand for part \( k \). The service coverage constraints state that the fraction of all failures that can be addressed from facilities within the time window is at least \( \alpha_k \) for part \( k \). We then write the service coverage constraints as follows:

\[
\sum_{i \in I} d_{ik}^{\leq w} d_{ik} / d_k \geq \alpha_k. \tag{3}
\]

The multi-part, service-coverage constrained, uncapacitated facility location model is as follows:

\[
\begin{align*}
\text{min} & \sum_{i \in I} f_i Y_i + \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} c_{ijk} d_{jk} X_{ijk} \\
\text{subject to} & \sum_{i \in I} X_{ijk} = 1, \ \forall j \in J, k \in K \\
& X_{ijk} \leq Y_i, \ \forall i \in I, j \in J, k \in K \\
& \sum_{i \in I} \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} \geq \alpha_k, \ \forall k \in K \\
& 0 \leq X_{ijk} \leq 1, \ \forall i \in I, j \in J, k \in K \\
& Y_i = 0 \text{ or } 1, \ \forall i \in I
\end{align*}
\]
The objective (4) is to minimize the total expected fixed facility opening costs and transportation costs. Constraints (5) guarantee all demands at all demand points to be fully satisfied. Constraints (6) mean that any facility serving a demand point should be open. As rewritten versions of (3) to explicitly show the allocation decision variables $X_{ijk}$, constraints (7) ensure that the required percentage of the total demand for each part is assigned to a facility within the specified time window of the demand points.

The implicit assumption made with this constraint is that the assignment of demands to facilities within the service time window guarantees delivering the part within the window. However, the actual service is not only a function of the transportation time (or distance) between the customers and their assigned facilities, but also a function of part availability. More interestingly, the part availability at a facility is a function of the demand assigned to it, which is just being decided with variables $X_{ijk}$ for all $i, j, k$: For a given stock level at the facility, as we assign more demand to a facility, its part availability deteriorates. The part availability is usually captured as long run fill rate, which is defined as the long run fraction of demand satisfied directly from stock on hand.

We now define the mean lead time demand at facility $i$ for part $k$:

$$\lambda_{ik} = t_{ik}d_{ik} = t_{ik}\sum_{j\in J} d_{jk}X_{ijk}$$ (10)

where $t_{ik}$ is the replenishment lead time for part $k$ at facility $i$, measured in years. Hence, due to the assumptions we made earlier, the lead time demand for part $k$ at facility $i$ is Poisson with mean rate $\lambda_{ik}$. Also, let $S_{ik}$ be the base stock level for part $k$ at facility $i$. We can now define $\beta_{ik}(S_{ik}, \lambda_{ik})$ as the fill rate for part $k$ at facility $i$ with (base) stock level $S_{ik}$ and mean lead time demand $\lambda_{ik}$. As demands (part failures) occur one at a time according to a Poisson process and the demands that cannot be filled immediately from stock are backordered, from the PASTA (Poisson Arrivals See Time Averages) property, the fill rate is equal to the limiting probability of having stock on hand (Zipkin, 2000; Muckstadt, 2005). Then, for a given mean lead time demand and a stock level for a part at a facility, the fill rate is

$$\beta_{ik}(S_{ik}, \lambda_{ik}) = G_{ik}(S_{ik} - 1) = \sum_{r=0}^{S_{ik}-1} \frac{\lambda_{ik}^r e^{-\lambda_{ik}}}{r!}$$ (11)

where $G_{ik}(S_{ik} - 1)$ is the cumulative (Poisson) distribution function of the lead time demand for part $k$ at facility $i$ evaluated at $S_{ik} - 1$. Note that the dependency of fill rate and lead time demand (which in turn is a function of demand allocation decisions, $X_{ijk}$’s) is implicit in the cumulative distribution function. An illustration of fill rate as a function of the mean lead time demand and the stock level is shown in Figure 1 (depicted for two stock levels).
With the fill rates calculated correctly, the more accurate service levels that include the possibility that the parts may not be available at the facilities when needed can now be written as follows:

\[ \sum_{i \in I} d_{ik}^w \frac{d_{ik}}{d_k} \beta_k(S_{ik}, \lambda_{ik}) \geq \alpha_k, \]  

(12)

or with the original demand allocation decision variables:

\[ \sum_{i \in I} \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} \frac{d_{ik}}{d_k} \beta_k(S_{ik}, \lambda_{ik}) \geq \alpha_k. \]  

(13)

Note that this constraint is similar in spirit to the service level constraints developed in (Muckstadt, 2005) for a given set of stocking facilities and their already allocated Poisson demands. In this new form needed in the integrated network design and stocking framework, the constraint is much more complex (and nonlinear), but is indeed critical for capturing the correct relationships among all elements of the overall system; facilities, transportation, inventory, and service.

We now write the complete model. To facilitate the model, we assume that there is a finite number of alternatives we consider for stock levels \( S_{ik} \). We denote this set of all possible stock levels by \( L \), and denote its largest element by \( L \). Hence, we consider \( S_{ik} \in \{0, 1, 2, \ldots, L\} \) for all \( i, k \). For example, typical stock levels at low-echelon facilities in most SPL systems are usually 0, 1, sometimes 2, and very rarely 3. As SPL systems deal with extremely low demand rates, considering higher stock levels is unnecessary and treating stock levels continuous is not an option. As the main idea here is to capture the fill rates for changing stock levels and compute its inventory costs, we make this set large enough to give practically 100% fill rate at the largest stock level \( L \) for a conservative estimate of total demand that could be assigned to a facility. In this case, even if we included larger stock levels in the set, the model would not consider them as any additional unit would increase inventory costs without providing additional improvement in fill rate and service. Finally, let \( h_{ik} \) be the unit inventory holding cost of part \( k \) at facility \( i \). The overall model is as follows:

\[
\begin{align*}
\min & \sum_{i \in I} f_i Y_i + \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} c_{ijk} d_{jk} X_{ijk} + \sum_{i \in I} \sum_{k \in K} h_{ik} S_{ik} \\
\text{subject to} & \sum_{i \in I} X_{ijk} = 1, \forall j, k \tag{14} \\
& X_{ijk} \leq Y_i \forall i, j, k \tag{15} \\
& S_{ik} \leq L Y_i, \forall i, k \tag{16} \\
& \lambda_{ik} = t_{ik} \sum_{j \in J} d_{jk} X_{ijk}, \forall i, k \tag{17} \\
& \beta_{ik}(S_{ik}, \lambda_{ik}) = S_{ik}^{r-1} \frac{e^{-\lambda_{ik}}}{r!}, \forall i, k \tag{18}
\end{align*}
\]
\[ 0 \leq X_{ijk} \leq 1, \forall i \in I, j \in J, k \in K \]  
\[ Y_i = 0 \text{ or } 1, \forall i \in I \]  
\[ S_{ik} \in \{0, 1, 2, \ldots, L\}, \forall i, k \] 

The new term in the objective (14) is the total inventory investment. Constraints (17) allow stock levels to be greater than 0 only for open facilities. Constraints (18) are the time-based service coverage constraints, where the mean lead time demand rates are calculated in constraints (19) and the fill rates in constraints (20). Finally, constraints (23) allow the stock levels to be selected from the initially developed set of integer stock levels.

This model is clearly a highly nonlinear, mixed integer programming problem. The service level constraints (18) are the main source of nonlinearity, since they combine demand allocation variables \( X_{ijk} \)'s with the fill rate variables \( \beta_{ik}(S_{ik}, \lambda_{ik}) \), which are themselves nonlinear functions of two other decision variables, stock levels \( S_{ik} \) and mean lead time demands \( \lambda_{ik} \), which are in turn functions of demand allocation variables \( X_{ijk} \)'s. Linearization of the service level constraints is the topic of the next subsection.

### 3.3 Linearized Model

We first approximate the nonlinear fill rate function with a step function. The motivation behind this development is the possibility that we can tabulate potential fill rates for a set of demand levels and stock levels \( a \text{ priori} \), and turn complex fill rate and mean lead time calculations into table-lookups. We then separate the nonlinearity using additional variables and finally remove the nonlinearity in the service level constraints.

We define the following additional notation to detail our approximation: Suppose we divide the mean lead time demand axis into \( N \) intervals. The intervals are indexed by \( n \) and the right end point of each interval is denoted by \( a_{kn} \) (these are break points on the demand axis for each part, where \( a_{k0} = 0 \)). Hence, when the mean lead time demand falls in the interval \( (a_{k,n-1}, a_{kn}] \), its fill rate for a stock level is approximated as the (actual) fill rate of the mid-point of the interval evaluated at the same stock level. More specifically, let \( b_{kl} \) be the approximated fill rate when the mean lead time demand is in \( (a_{k,n-1}, a_{kn}] \) and when the stock level under consideration is \( l \). Then,

\[ b_{kl} = G(l - 1) \]  

where \( G(.) \) now is the Poisson cumulative distribution function with the mean lead time demand of \( (a_{k,n-1} + a_{kn})/2 \). The use of mid-point mean lead time demand aims to prevent constant over- or under-estimation of the true fill rates. (Later, we outline a post-processing phase that will compute actual fill rates and revise stock levels when necessary.) Finally, we do a similar approximation for all potential stock levels, \( l = 1, 2, \ldots, L \).
We note that the demand intervals are not necessarily of equal size; in fact we intentionally make the intervals larger as the mean lead time demand increases (as illustrated in Figure 1). The idea is to capture the sensitive and useful parts of the fill rate function more accurately (with a higher resolution) in the parts of the demand most relevant to our calculations. Due to low demand levels, we are more likely and more heavily to use the left-tail of the fill rate function than the approximations on the right-tail (larger demand values), which may not be used at all.

![Figure 1: Fill rate linearization, shown for two stock levels](image)

We now formulate the table-lookup process, finding the correct demand interval in which the mean lead time demand falls in, and making this process part of the model. For this purpose, we define another binary variable, \(Q_{ikn}\), which takes value 1 when \(a_{k,n-1} < \lambda_{ik} \leq a_{kn}\), and takes value 0 otherwise. We write the constraints that guarantee these assignments with the help of another set of variables, \(R_{ikn}\), defined for all \(i, k, n\).

\[
a_{kn} - \lambda_{ik} \geq M_1(R_{ikn} - 1), \forall i, k, n \tag{25}
\]

\[
a_{kn} - \lambda_{ik} \leq M_2 R_{ikn}, \forall i, k, n \tag{26}
\]

\[
Q_{ikn} = R_{i,k,n} - R_{i,k,n-1}, \forall i, k, n \tag{27}
\]

\[
Q_{ikn} = 0 \text{ or } 1, R_{ikn} = 0 \text{ or } 1, \forall i, k, n \tag{28}
\]

\[
R_{ik0} = 0, \forall i, k \tag{29}
\]

where \(M_1\) and \(M_2\) are big-\(M\)’s that can be set to their tightest values, desirably depending on the facility, part, and demand interval: \(M_1 = t_{ik}d_k\) (the largest possible value of the mean lead time demand for part \(k\) at facility \(i\)) and \(M_2 = a_{kn}\).
Here $R_{ikn} = 0$ for $n$ values, up to the interval just before the demand interval that the mean lead time demand $\lambda_{ik}(X_{ik})$ falls in. For this and other intervals $R_{ikn} = 1$. Hence, $Q_{ikn} = 1$ for the interval that has the mean lead time demand in it. This is where $R_{i,k,n-1} = 0$ and $R_{ikn} = 1$. We now write the following for the approximate fill rates for given stock level $S_{ik} = l$:

$$\beta_{ik}(S_{ik} = l, \lambda_{ik}) = \sum_{n \in N} b_{kln} Q_{ikn}, \forall i, k$$

where $N$ is the set of numbers from 0 to the number of intervals used in the fill rate approximation.

To make further linearization possible for a generic $S_{ik}$, we define a new set of decision variables: $W_{ikl}$ takes value 1 when facility $i$ uses stock level $l$ for part $k$, and 0 otherwise. With this, we write the approximate fill rates as follows:

$$\beta_{ik}(S_{ik}, \lambda_{ik}) = \sum_{l \in L} \sum_{N_{b_{kln}}} W_{ikl} Q_{ikn}, \forall i, k$$

where $W_{ikl}$ will be forced to take value of 1 for $l = S_{ik}$, 0 otherwise. Note that this final version of fill rate approximation consists of multiplication of two binary variables ($W_{ikl}$ and $Q_{ikn}$). We can plug in these versions of approximate fill rates into the service level constraints (18) to obtain the following (while eliminating constraints (20)).

$$\sum_{i \in I} \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} \left( \sum_{l \in L} \sum_{n \in N} b_{kln} W_{ikl} Q_{ikn} \right) \geq \alpha_k, \forall k \in K.$$  (32)

Constraint (32) for part $k$ simply states that for “selected stock level $l$” (hence $W_{ikl} = 1$), for the “corresponding mean lead time demand” ($Q_{ikn} = 1$), and for the approximated fill rates for each facility $i$, the sum of “fractions of demands that are directly satisfied from facilities that are within the time window” ($\sum_{i \in I} \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} b_{kln}$) should be at least the “target service level” ($\alpha_k$).

To finalize our linearization, we define a new variable, $V_{ikln}$ which takes the value $\sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} b_{kln}$ for $W_{ikl} = Q_{ikn} = 1$, and 0 otherwise. Hence, the service level satisfied at facility $i$ for part $k$ is given by $\sum_{l \in L} \sum_{n \in N} V_{ikln}$. To achieve this, we introduce the following constraints:

$$V_{ikln} \leq M_3 Q_{ikn}, \forall i, k, l, n$$  (33)

$$V_{ikln} \leq M_3 W_{ikl}, \forall i, k, l, n$$  (34)

$$V_{ikln} \leq \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} b_{kln}, \forall i, k, l, n$$  (35)

$$V_{ikln} \geq \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} b_{kln} - M_3 (1 - Q_{ikn}) - M_3 (1 - W_{ikl}), \forall i, k, l, n$$  (36)

$$V_{ikln} \geq 0, \forall i, k, l, n$$  (37)

where $M_3$ is another big-$M$ with the tightest value that can be customized for each $i, k, l, and n$: $M_3 = \sum_{j \in J} \delta_{ij} d_{jk} b_{kln}$. Here, if one or both of $W_{ikl}$ and $Q_{ikn}$ is 0 for some $i, k, l$ and $n$, then
constraints (33, 34, 37) force $V_{ikln}$ to be 0. For part $k$ at facility $i$, if both $W_{ikl}$ and $Q_{ikn}$ are 1 for some $l$ and $n$, i.e., say $\hat{l}$ and $\hat{n}$ ($W_{ik\hat{l}} = 1$ and $Q_{ik\hat{n}} = 1$), constraints (35) and (36) will force $V_{ik\hat{l}\hat{n}}$ to be equal to $\sum_{j \in J} \frac{\delta_{ij}d_{ijk}X_{ijk}}{d_k} b_{k\hat{l}\hat{n}}$. We can now rewrite the service level constraints (32) as follows:

$$\sum_{i \in I} \sum_{l \in L} \sum_{n \in N} V_{ikln} \geq \alpha_k, \forall k$$ (38)

The following is the complete model:

$$\min \sum_{i \in I} f_i Y_i + \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} c_{ijk} d_{ijk} X_{ijk} + \sum_{i \in I} \sum_{k \in K} \sum_{l \in L} l_{ik} W_{ikl}$$ (39)

subject to

$$\sum_{i \in I} X_{ijk} = 1, \forall j, k$$ (40)

$$X_{ijk} \leq \sum_{l \in L} W_{ikl}, \forall i, j, k$$ (41)

$$\sum_{l \in L} W_{ikl} \leq Y_i, \forall i, k$$ (42)

$$\lambda_{ik} = t_{ik} \sum_{j \in J} d_{ijk} X_{ijk}, \forall i, k$$ (43)

$$R_{ik0} = 0, \forall i, k$$ (44)

$$a_{kn} - \lambda_{ik} \geq M_1 (R_{ikn} - 1), \forall i, k, n$$ (45)

$$a_{kn} - \lambda_{ik} \leq M_2 R_{ikn}, \forall i, k, n$$ (46)

$$Q_{ikn} = R_{i,k,n} - R_{i,k,n-1}, \forall i, k, n$$ (47)

$$V_{ikln} \leq M_3 Q_{ikn}, \forall i, k, l, n$$ (48)

$$V_{ikln} \leq M_3 W_{ikl}, \forall i, k, l, n$$ (49)

$$V_{ikln} \geq 0 \forall i, k, l, n$$ (50)

$$V_{ikln} \leq \sum_{j \in J} \frac{\delta_{ij}d_{ijk}X_{ijk}}{d_k} b_{kln}, \forall i, k, l, n$$ (51)

$$V_{ikln} \geq \sum_{j \in J} \frac{\delta_{ij}d_{ijk}X_{ijk}}{d_k} b_{kln} - M_3 (1 - Q_{ikn}) - M_3 (1 - W_{ikl}), \forall i, k, l, n$$ (52)

$$\sum_{i \in I} \sum_{l \in L} \sum_{n \in N} V_{ikln} \geq \alpha_k, \forall k$$ (53)

$$Y_i, W_{ikl}, Q_{ikn}, R_{ikn} = 0 \text{ or } 1, \forall i, k, l, n$$ (54)

The formulation has about $I(1+K(L+2N+LN+J)$ variables, and $K+JK+IK(J+3+3N+5LN)$ constraints. For example, with 10 facilities, 20 customers, 2 parts, 5 stock levels, and 5 fill rate function steps, we have 1210 variables and 3302 constraints. For a problem instance with 25
facilities, 150 customers, 10 parts, 5 stock levels, and 10 fill rate function steps, we have 55,025
variables and 106,010 constraints.

This rather large, but linear, integer programming model captures all the cost tradeoffs and the
complex relationships among different entities of the overall problem. Taking advantage of the
low-demand nature of the parts, the linearization process makes this possible, but it comes with
a cost: Fill rate approximation used in the process is indeed an approximation, hence when the
demands are allocated and stock levels are computed, the actual fill rates may be different from the
fill rates approximated in the model. In fact, there are two possibilities: (1) There is potential for
an overestimation of fill rate in the model, in which the solution may not actually satisfy the service
level constraints (hence, may become infeasible) when the actual fill rates are considered. (2) There
is also potential for an underestimation in the model, in which the solution may be overstocking,
increasing the costs unnecessarily. To relieve these issues, we add a post processing stage.

3.4 Post Processing

We formalize the process of uncovering the potentials for revising stock levels with a post processing
stage. This post processing stage will revise the stock levels, if necessary, (1) to guarantee ultimate
feasibility of the solution (for service constraints) with minimal cost increase, and (2) to save in
inventory costs as much as possible without sacrificing from actual service (that is, still satisfying
the required service level). To do this, we take the location and allocation decisions of the solution
of the linearized model as given, and solve a multi-location inventory model to revise the stock level
decisions (only if necessary) at the open facilities. As expected, this model borrows constraints from
the linearized model, and is significantly smaller and simpler.

Suppose we solve the linearized model (39-54), and obtain the solution with location decisions
\( Y_i = \hat{Y}_i \) for all \( i \), demand allocation decisions \( X_{ijk} = \hat{X}_{ijk} \) for all \( i, j \), and \( k \). As we focus on the
open facilities only at this stage, we define \( \hat{I} \) to be the set of open facilities. Once these decisions
are known (and fixed), we compute the mean lead time demand for each part at the open facilities:

\[
\hat{\lambda}_{ik} = t_{ik} \sum_{j \in J} d_{jk} \hat{X}_{ijk}, \quad \forall i \in \hat{I}
\]  

Moreover, we can now compute the actual fill rates:

\[
\hat{\beta}_{ik}(l) = G_{ik}(l-1) = \sum_{r=0}^{l-1} \frac{(\hat{\lambda}_{ik})^r e^{\hat{\lambda}_{ik}}}{r!}
\]

where \( G_{ik}(\cdot) \) is the Poisson distribution function with mean lead time demand \( \hat{\lambda}_{ik} \) for part \( k \)
calculated at each stock level \( l \), \( l = 0, 1, 2, \ldots, L \), at the open facilities \( i \in \hat{I} \). We then use these fill
rates to rewrite the service level constraints, and decide if we need to revise the stocking decisions.
The following is a listing of the multi-location inventory model, whose only set of decision variables is the stock levels, that are modeled using the previously defined variables $W_{ikl}$’s.

$$\min \sum_{i \in \hat{I}} \sum_{k \in K} \sum_{l \in L} l h_{ik} W_{ikl}$$ (57)

$$\sum_{l \in L} W_{ikl} \geq \hat{X}_{ijk}, \forall i \in \hat{I}, j \in J, k \in K$$ (58)

$$\sum_{i \in \hat{I}} \sum_{j \in J} \sum_{l \in L} \frac{\delta_{ij} d_{jk}}{d_k} \beta_{ik}(l) W_{ikl} \geq \alpha_k \forall k \in K$$ (59)

$$W_{ikl} = 0 \text{ or } 1, \forall i \in \hat{I}, k \in K, l \in L$$ (60)

The objective (57) of the multi-location inventory model is the minimization of the total inventory costs, as the network design decisions are kept the same. Constraints (58) are forcing constraints to keep a part’s stock at the open facilities serving that part. Service level constraints (59) are much simpler, due to the fixed network design decisions. The integrated approach solution is then the combination of the network design (location and demand allocation) decisions from the linearized formulation and the inventory stocking decisions from the multi-location inventory formulation (if different from the linearized model’s solution). The total cost can be calculated accordingly.

4 Decoupled Approach

A goal of this study is to show the benefit of integrated modeling of network design and inventory stocking. We now introduce the “decoupled model” that we use to emulate the conventional and practical approach of sequentially designing the network first and choosing the stock levels later.

In fact, the decoupled model actually consists of two submodels: (1) Logistics network design submodel (LND-only submodel), and (2) Inventory stocking submodel (IS-only submodel). LND-only submodel locates the facilities of the network and allocates demands to these facilities without considering inventory costs and assuming practically free 100% fill rate for all parts at all (open) facilities. IS-only submodel takes the network and the demand allocations as input and decides the stock levels for all facilities and parts while trying to satisfy time-based service levels. Both submodels borrow notation from the integrated model. In fact, separation of the two in the decoupled model simplifies the service constraints, and makes both submodels integer linear programming problems. Hence, we list the submodels without defining new notation or decision variables.

The LND-only submodel is a version of the service-constrained, multi-commodity uncapacitated facility location problem, which is practically the same as the one used to introduce the integrated modeling earlier in Section 3.2:

$$\min \sum_{i \in \hat{I}} f_i Y_i + \sum_{i \in \hat{I}} \sum_{j \in J} \sum_{k \in K} c_{ijk} d_{jk} X_{ijk}$$ (61)
\[
\sum_{i \in I} X_{ijk} = 1, \quad \forall j \in J, k \in K \quad (62)
\]
\[
X_{ijk} \leq Y_i, \quad \forall i \in I, j \in J, k \in K \quad (63)
\]
\[
\sum_{i \in I} \sum_{j \in J} \delta_{ij} d_{jk} X_{ijk} \geq \alpha_k, \quad \forall k \in K \quad (64)
\]
\[
0 \leq X_{ijk} \leq 1, \quad \forall i \in I, j \in J, k \in K \quad (65)
\]
\[
Y_i = 0 \text{ or } 1, \quad \forall i \in I \quad (66)
\]

All the objective function terms and constraints are self explanatory. From the optimal solution of the LND-only submodel, we obtain \( \tilde{I} \) as the set of open facilities (\( \tilde{I} = \{ i \in I : Y_i = 1 \} \)) and the demand allocation decisions \( \tilde{X}_{ijk} \) for all \( i, j, \) and \( k \). Using these, we calculate the actual mean lead time demands for all open facilities and parts:

\[
\tilde{\lambda}_{ik} = t_{ik} \sum_j d_{jk} \tilde{X}_{ijk}, \quad \forall i \in \tilde{I}, k \in K. \quad (67)
\]

We then compute the actual fill rates for all parts and facilities, for each potential stock level:

\[
\tilde{\beta}_{ik}(l) = G_{ik}(l - 1) = \frac{(\tilde{\lambda}_{ik})^{l-1} e^{\tilde{\lambda}_{ik}}}{(l-1)!}. \quad (68)
\]

These all become input to the IS-only submodel, which is itself a multi-facility inventory model, where the only decision variables are the stock levels, \( W_{ikl} \)'s. The submodel finds the minimum-cost stock levels possible to achieve service levels using the actual fill rates. As the network and demand allocations are decided and fixed already, the remaining objective is to minimize the total inventory costs. Thus, IS-only submodel is similar to the post-processing model of the integrated approach:

\[
\min \sum_{i \in \tilde{I}} \sum_{k \in K} \sum_{l \in L} h_{ik} W_{ikl} \quad (69)
\]
\[
\sum_{l \in L} W_{ikl} \geq \tilde{X}_{ijk}, \quad \forall i \in \tilde{I}, j \in J, k \in K \quad (70)
\]
\[
\sum_{i \in \tilde{I}} \sum_{j \in J} \sum_{l \in L} \delta_{ij} d_{jk} \tilde{X}_{ijk} \beta_{ik}(l) W_{ikl} \geq \alpha_k \forall k \in K \quad (71)
\]
\[
W_{ikl} = 0 \text{ or } 1, \quad \forall i \in \tilde{I}, k \in K, l \in L \quad (72)
\]

Combining the network from the LND-only submodel and the stocking decisions from the IS-only submodel, the total cost of the decoupled approach is the sum of the submodels’ two objectives.

### 5 Computational Study

#### 5.1 Experimental Data Set

We take our industrial partner’s real data as basis for our problem instances. The original data set has different sizes of problems up to 20 candidate facilities, 150 customers, and 10 parts. Here, we report the results of the problem instances for the following data set:
• **Candidate facilities** \((I)\) and **customers** \((J)\): We use the problem instances with 16 candidate facilities and 134 customers which are given by their zip codes in the real data.

• **Parts** \((K)\) and **part demands** \((d_{jk})\): There are 4 parts with different demand patterns. Part mean demands are given in the real data. To isolate the effects of different factors, we experiment with the single-part instances first by analyzing each demand pattern separately. Later, we show results from experiments with all 4 parts considered at the same time.

• **Fixed facility costs** \((f_i)\), **inventory holding costs** \((h_{ik})\), and **transportation costs** \((c_{ijk})\): We generate these costs randomly from discrete uniform distributions and scale them with cost factors to control their relative weights in the objective function. Fixed costs are generated as:

\[
    f_i \sim \theta_F U[500, 1500]
\]

where \(\theta_F\) is the *fixed facility cost multiplier* that is modified in the experiments. We use three different values for \(\theta_F\): 1, 10, and 100. Note that average fixed cost is 1000 when \(\theta_F = 1\), and 100,000 when \(\theta_F = 100\). Similarly, we calculate the holding costs using

\[
    h_{ik} \sim \theta_H U[0.25 \times v_k/2, 0.25 \times 3v_k/2]
\]

where \(\theta_H\) is the inventory holding cost multiplier to be modified to create different experimental settings, and \(v_k\) is the part \(k\)’s unit cost (modified from the real data). We use two different values for \(\theta_H\): 1, and 10. Transportation costs are given in the real data, which are loosely based on distances between facilities and customers, and sometimes on part specifications. We scale these values with the transportation cost multiplier \(\theta_T\) which takes either 1 or 10.

The cases with low fixed facility costs and/or high holding costs may seem unrealistic, but we note these are instances with up to 4 parts where inventory holding costs can be magnified only through their high holding costs. Most real SPL systems service thousands of high-cost parts, added together becoming a significant part of the overall cost. We tried to achieve a level of realism with similar compositions of the total costs by changing the cost multipliers. Overall, we generate 3 random instances for each setting of cost multipliers \(\theta_F, \theta_T, \text{and} \theta_H\).

• **Time windows** \((w)\): We use two different settings for time windows: 2 hours and 4 hours, which are typical in real SPL systems. Time window identifier \(\delta_{ij}\) for each facility and customer pair is given in the real data.

• **Lead times** \((t_{ik})\): A lead time of one week for all facilities and parts, as in the real data.

• **Maximum stock level** \((L)\): The maximum possible mean lead time demand (calculated from real data assuming all customer demands for a part are assigned to one virtual location) is quite variable, changing between 0.2 and 0.8 across parts. Hence, we use a maximum stock level \((L)\) of 5, at which the fill rate for even a virtual facility serving all possible demand is practically 100%.

• **Fill rate approximation** \((\beta_{ik})\): To facilitate the linearization of fill rates for a part, we divide the demand axis (0 to maximum possible lead time demand for the part) into \(N = 10\) intervals.
Table 1: Experimental design

<table>
<thead>
<tr>
<th>Factor</th>
<th>Description</th>
<th>Number of levels</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_F$</td>
<td>facility fixed cost multiplier</td>
<td>3</td>
<td>1, 10, 100</td>
</tr>
<tr>
<td>$\theta_H$</td>
<td>inventory holding cost multiplier</td>
<td>2</td>
<td>1, 10</td>
</tr>
<tr>
<td>$\theta_T$</td>
<td>transportation cost multiplier</td>
<td>2</td>
<td>1, 10</td>
</tr>
<tr>
<td>$w$</td>
<td>time windows</td>
<td>2</td>
<td>2 and 4 hours</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>target service levels</td>
<td>4</td>
<td>0.1, 0.3, 0.5, 0.7</td>
</tr>
<tr>
<td>$K$</td>
<td>demand patterns</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>number of instances</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Total number of instances: 1152

Since the most relevant demand values for actual facilities will be in the region with very low lead time demand, we use smaller intervals in that area, making wider intervals as demand increases.

- **Time-based service levels** ($\alpha_k$): We use the same target service level for all parts, varying them in the experiments: (10%, 30%, 50% and 70%).

We test our model with a wide variety of settings, changing cost coefficients, demand patterns and service requirements as shown in Table 1. In total, we solve 1152 problem instances. We use Xpress-MP branch-and-bound based solver (by Dash Optimization) to solve these instances.

5.2 Lower Bounds and Quality of Fill Rate Approximation

Because even small problem instances are difficult to solve to optimality, the only reasonable ways to measure the quality of our fill rate approximation and its associated solution are (1) to find a lower bound solution and measure the gap between the lower bound and the reported solution of the approximated model, and (2) to increase the accuracy of the fill rate approximation and show how much improvement could be obtained with the increased accuracy. To do (1), we use an overestimating fill rate approximation in the linearized model, i.e. define the step-function based approximation with (optimistic) maximum points at each step (not with mid-points of the intervals of the original fill rate function as in Figure 1). In a way, we draw the approximating step function above the actual fill rate curve. To do (2), we solve the linearized model with two more levels of the number of steps in the approximation ($N = 5$ and $N = 15$), along with $N = 10$ steps.

Table 2 reports the linearized model costs for different levels of fill rate approximations (with 5 steps, 10 steps and 15 steps), and the percentage gaps between the tightest lower bound solution and the integrated model solution (obtained with 10-step fill rate approximation). The experiments are conducted for one demand pattern (A), one instance (data set 1), with 4-hour service time window, and 70% service level requirement. We limit the solution times to two hours, and the instances shown with * are not finished within this time limit (hence not guaranteed to be optimal).
As expected, as we increase the accuracy of the fill rate approximation, we obtain better solutions from the linearized model. The 10-step approximation seems to provide the best compromise between the solution quality and computational time. Moreover, the reported solutions of the 10-step approximation are mostly within 1% of the lower bound, signaling the fact that the solutions we provide are near-optimal. Hence, we use the 10-step approximation in our experiments.

### 5.3 Integrated versus Decoupled Approaches

In this section, first, we show how much one saves in total costs by using the integrated approach instead of the decoupled approach, and investigate how these savings change with different problem settings and conditions. We then compare their actual solutions, namely, resulting networks (number and locations of open facilities, demand allocations and inventory decisions) to better understand the sources of the savings.

Figure 2 shows the histograms of the saving percentages of the integrated approach over the decoupled approach for each specific set of cost coefficients. Saving percentage is calculated as the total cost difference between the integrated approach and the decoupled approach divided by the decoupled approach’s total cost multiplied by 100.

The integrated approach achieves arguably significant cost savings when the fixed facility and transportation costs are relatively low and/or the holding costs are relatively high. As in Figure 3.c, there are some instances (almost 10 of them) achieving cost savings of more than 80% for $\theta_H = 10$. Interpreted in another way, the decoupled approach emulating the current practice of sequentially deciding the network and then the stock levels may cost up to 5 times the cost of making the decisions concurrently. The savings are lower when the opposite conditions are present.

When the facility and transportation costs are high and holding costs are low, the overall problem resembles the uncapacitated service-constrained facility location problem, which means inventory stocking loses its importance. In these cases, the decoupled approach produces quite accurate decisions in the LND-only submodel while ignoring inventory. We can see this behavior of the models in Figure 3, in which we fix the facility and transportation costs to various combinations of the multipliers’ levels. Each graph shows the histogram for both low and high holding costs. From these, we can easily state that the decoupled approach fails more often and more significantly when inventory costs are larger.

<table>
<thead>
<tr>
<th>$\theta_H$</th>
<th>$\theta_T$</th>
<th>$\theta_F$</th>
<th>$\theta_H$</th>
<th>$\theta_T$</th>
<th>$\theta_F$</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>100</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>$N = 5$</td>
<td>3979</td>
<td>16251</td>
<td>133521</td>
<td>11841</td>
<td>26266</td>
<td>143556</td>
</tr>
<tr>
<td>$N = 10$</td>
<td>3841</td>
<td>15604</td>
<td>132874</td>
<td>11415</td>
<td>26263</td>
<td>143590</td>
</tr>
<tr>
<td>$N = 15$</td>
<td>3833</td>
<td>15602</td>
<td>132872</td>
<td>11816*</td>
<td>26248</td>
<td>143548</td>
</tr>
<tr>
<td>LB</td>
<td>3835</td>
<td>15599</td>
<td>132869</td>
<td>11305</td>
<td>25781</td>
<td>143517</td>
</tr>
<tr>
<td>% Gap</td>
<td>0.14</td>
<td>0.03</td>
<td>0.00</td>
<td>0.97</td>
<td>1.87</td>
<td>0.05</td>
</tr>
</tbody>
</table>

As expected, as we increase the accuracy of the fill rate approximation, we obtain better solutions from the linearized model. The 10-step approximation seems to provide the best compromise between the solution quality and computational time. Moreover, the reported solutions of the 10-step approximation are mostly within 1% of the lower bound, signaling the fact that the solutions we provide are near-optimal. Hence, we use the 10-step approximation in our experiments.

<table>
<thead>
<tr>
<th>$\theta_H$</th>
<th>$\theta_T$</th>
<th>$\theta_F$</th>
<th>$\theta_H$</th>
<th>$\theta_T$</th>
<th>$\theta_F$</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>100</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>$N = 5$</td>
<td>3979</td>
<td>16251</td>
<td>133521</td>
<td>11841</td>
<td>26266</td>
<td>143556</td>
</tr>
<tr>
<td>$N = 10$</td>
<td>3841</td>
<td>15604</td>
<td>132874</td>
<td>11415</td>
<td>26263</td>
<td>143590</td>
</tr>
<tr>
<td>$N = 15$</td>
<td>3833</td>
<td>15602</td>
<td>132872</td>
<td>11816*</td>
<td>26248</td>
<td>143548</td>
</tr>
<tr>
<td>LB</td>
<td>3835</td>
<td>15599</td>
<td>132869</td>
<td>11305</td>
<td>25781</td>
<td>143517</td>
</tr>
<tr>
<td>% Gap</td>
<td>0.14</td>
<td>0.03</td>
<td>0.00</td>
<td>0.97</td>
<td>1.87</td>
<td>0.05</td>
</tr>
</tbody>
</table>
As shown in Figure 4, the integrated approach yields more savings for instances with 4-hr time window than those with 2-hr time window. When time windows are longer, the decoupled approach tends to assign more demand to each facility, potentially using fewer facilities (depending on the fixed facility costs), without considering inventory implications of these decisions. Using the LND-only solution, the IS-only submodel tries to achieve very high fill rates for each of these facilities with high demands to be able to satisfy the service levels. With more demand assigned per facility, IS-only submodel can do this only by keeping high stock levels, causing the total costs to increase significantly. The integrated approach in these settings is able consider all these interactions by design and finds the best tradeoff between all cost components and service constraints.

As summary of the single-part results, Figure 5 shows the tradeoff curves (also called the “efficient
Figure 3: Cost savings obtained via the integrated approach over the decoupled approach for varying levels of the cost multipliers (two levels of fixed facility and two levels of transportation costs), where $\theta_H$ is varied in each plot.

(a) Fixed levels of $\theta_F = 1, \theta_T = 10$
(b) Fixed levels of $\theta_F = 100, \theta_T = 10$
(c) Fixed levels of $\theta_F = 1, \theta_T = 1$
(d) Fixed levels of $\theta_F = 100, \theta_T = 1$

The figure shows that one can shift the whole tradeoff curve to the right (superior regions) by integrating network design and stocking. For example, in order to achieve 50% service level the decoupled approach suggests that we need almost $124,000 on average (see Figure 5.c). However,
the integrated approach requires only $108,000 to guarantee the same level of service. Similarly, with $124,000 we can achieve 50% service level using the decoupled approach, while we can achieve more than 70% service level with the integrated approach with the same budget. As shown in the figure, the savings become larger for higher target service levels, achieving arguably significant cost savings at certain settings such as that of Figure 5.a.

We now investigate under what conditions the solutions of the two approaches differ, eventually producing sometimes significant total costs differences. Table 3 shows the network differences between the integrated and the decoupled optimal solutions. In this table,

- 0 means the two produce totally different solutions (different numbers of open facilities, different demand allocations, and different stock levels),
- 1 means the number of open facilities is the same in both solutions, but everything else is different,
- 2 means both the number and the locations of open facilities are the same, but the demand allocations and the stock levels at these facilities are different between the two solutions,
- 3 means all aspects of the solutions are the same except the stock levels, and finally
- 4 means that the two solutions are exactly the same.

The table provides a lot of detail about the solutions of the two approaches. For example, the two approaches produce pretty much the same solution for the setting where $\theta_F = 100$, $\theta_T = 10$, and $\theta_H = 1$. However, the integrated approach finds a completely different solution (including the number and the locations of open facilities) for $\theta_F = 1$, $\theta_T = 1$, and $\theta_H = 10$. Under these conditions, using a decoupled approach, and, as a result, designing a suboptimal network may be costly, as the wrong network decisions will undermine the system performance for a long time before a revision of the network is undertaken (and corrected).

We can ultimately calculate a measure of average similarity of solutions by taking an average of the values for each setting. For example, the average similarity within $\theta_F = 100$, $\theta_T = 10$, $\theta_H = 1$ is 2.81.
(a) Time window $w = 4$-hr, $\theta_F = \theta_T = \theta_H = 1$  
(b) Average of all instances with 4-hr windows  
(c) Average of all 1152 instances

Figure 5: Tradeoff curves for the single-part experiments (Each point is an average of total (optimal) costs all instances with given settings under each chart)

for both demand patterns 1 and 3, whereas the average similarity within $\theta_F = 1, \theta_T = 1, \theta_H = 10$ is 1.59 for demand pattern 1 and even lower (0.81) for demand pattern 3. This shows the benefit of including multiple demand patterns, as they affect locations, allocations, and stock levels, hence collecting more detailed information regarding the differences between the two approaches.

The original model intends to capture the multi-part case as the real SPL systems support and stock multiple parts. Hence, we now present results of the instances where we consider all 4 demand patterns at the same time, each representing an individual part. Instead of a similar detailed analysis of the multi-part cases, for brevity, we show the total costs vs. service level tradeoff curves for the multiple-part instances. Figure 6 draws tradeoff curves for both the decoupled and integrated approaches for 4-hr time window and by using data instance 1 (in total, we run 48 instances, each point on the graphs representing the average of 12 instances: 3 levels of fixed costs, 2 levels of transportation costs, and 2 levels of holding costs). The figure shows parallel results to
<table>
<thead>
<tr>
<th>Demand Patterns</th>
<th>( \theta_P = 1 )</th>
<th>( \theta_P = 10 )</th>
<th>( \theta_P = 1 )</th>
<th>( \theta_P = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \theta_H = 1 )</td>
<td>( \theta_H = 1 )</td>
<td>( \theta_H = 1 )</td>
<td>( \theta_H = 1 )</td>
</tr>
<tr>
<td>A 2hr</td>
<td>0.3 0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
<tr>
<td>4hr</td>
<td>0.3 0.4 0.3 0.4 0.3</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
<tr>
<td>B 2hr</td>
<td>0.3 0.2 0.1 0.3 0.1</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
<tr>
<td>4hr</td>
<td>0.3 0.2 0.1 0.3 0.1</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
<tr>
<td>C 2hr</td>
<td>0.3 0.2 0.1 0.3 0.1</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
<tr>
<td>4hr</td>
<td>0.3 0.2 0.1 0.3 0.1</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
<tr>
<td>D 2hr</td>
<td>0.3 0.2 0.1 0.3 0.1</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
<tr>
<td>4hr</td>
<td>0.3 0.2 0.1 0.3 0.1</td>
<td>0.2 0.1 0.2 0.2</td>
<td>0.1 0.2 0.4 0.2</td>
<td>0.2 0.1 0.2 0.2</td>
</tr>
</tbody>
</table>

Table 3: Network comparison between the integrated and decoupled approaches.
Table 4: Computation times for the integrated model, post-processing, and decoupled model

<table>
<thead>
<tr>
<th></th>
<th>Integrated Model</th>
<th>Post Processing</th>
<th>Decoupled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>215.80</td>
<td>0.046</td>
<td>0.3</td>
</tr>
<tr>
<td>Min</td>
<td>0.48</td>
<td>0.03</td>
<td>0.094</td>
</tr>
<tr>
<td>Max</td>
<td>6461.95</td>
<td>0.33</td>
<td>2.69</td>
</tr>
</tbody>
</table>

the earlier ones and we make similar observations here as before.

![Figure 6: Tradeoff curves for the multi-part instances (4-hr time window and data 1)](image)

Although our focus is on the quantification of benefits of considering inventory as part of the network design problem, we give an overall feel for its computational difficulty by listing computer times (in CPU seconds) in Table 4. The table shows that even for the relatively small problem instances considered in the paper, the integrated model is a computationally challenging problem that needs further analysis, scalable solution techniques, and powerful heuristics. This is a suggested idea for future research.

### 5.4 Decoupled Approach with Varying Fill Rates

In the original decoupled approach, we had assumed that each facility has 100% fill rate in the LND-only submodel and we had solved for the required inventory levels in the IS-only submodel. This was done to imitate the conventional approach of solving a coverage-based facility location allocation model as the first step and then deciding the inventory levels for the given network in the second step. To see if improved solutions are possible within the decoupled approach, we introduce a refined version which solves the first step assuming that the open facilities will have lower (but
Table 5: Integrated model versus decoupled approach with different fill rates at first step

<table>
<thead>
<tr>
<th>Fill Rate (%)</th>
<th>Decoupled approach</th>
<th>Integrated</th>
<th>% gap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>95</td>
<td>90</td>
</tr>
<tr>
<td>0.1</td>
<td>5437.2</td>
<td>5386.16</td>
<td>5798.18</td>
</tr>
<tr>
<td>0.3</td>
<td>9039.47</td>
<td>7450.63</td>
<td>7939.91</td>
</tr>
<tr>
<td>0.5</td>
<td>16731.3</td>
<td>16736.64</td>
<td>16736.64</td>
</tr>
</tbody>
</table>

still fixed) fill rates such as 95%.

Table 5 shows the average costs across 12 instances (demand patterns A, B, C, and D; instances 1, 2, and 3), all run with the cost coefficients $\theta_F = \theta_T = \theta_H = 1$; 4 hour service window, and varying service level requirements (10% - 70%). We run the decoupled approach with changing fill rate values used in the first step (as part of the coverage constraint in the LND-only model) from 100% to 80% in 5% increments. We compare these solutions with the integrated model solution, and show the gap between the best of the decoupled approach (obtained by picking the lowest cost among the costs of the different fill rates for each of the 12 instances) and the integrated approach.

As shown in the table, it is possible to improve the decoupled approach by trying different fill rates. However, it is still not possible to obtain superior solutions as compared to the integrated solutions. In fact, we have an increasing improvement obtained from the integrated approach as compared to the variable fill rate scheme for the decoupled approach. Especially at high service levels (50% and 70%), we have significant improvements by integration, up to 16% compared to the best decoupled approach. We show the best performing fill rates in bold font for each service level requirement. Note that, different fill rates yield the best performance for different service levels, which may mean that one has to run many fill rates in the decoupled approach to find an overall good solution.

6 Conclusions and Future Work

Our major conclusion is that we can achieve same service levels with less cost with the integrated approach when compared with the traditional approach of making these decisions sequentially. Similarly, higher service levels can be achieved for a given budget with the integrated approach. We gain increasingly significant benefits from integration when inventory decisions become more dominant (i.e., the problems with high inventory holding costs caused by expensive parts, higher inventory levels due to longer time windows, or higher required service levels). Since the decoupled approach inherently ignores the inventory decisions while designing the network, when these decisions become more dominant and interact with network design decisions more, the cost of ignoring them is magnified.
There are many ways the model can be extended to make it more comprehensive and realistic. We can extend the model to consider multiple regions, each with regional time-based service levels, and/or multi-tiered service time windows, which is useful when considering the systems with customers having products with different criticality levels. Other extensions for SPL-based problems can include part commonality across products in a multi-product setting, multi-echelon systems, and inventory sharing and transshipments among facilities. Each of these extensions will make the already challenging and rich problem introduced here even more interesting. Another avenue for future research is to develop more efficient solution techniques, e.g. those based on decomposition, column generation, and Lagrangian relaxation. Another possibility is to consider an iterative technique based on the idea of repeating the stages of the decoupled approach in a controlled manner.

7 Acknowledgements

This research is supported in part by by NSF CAREER Grant DMI-0245123. We thank the anonymous referees and the associate editor whose comments significantly improved the paper.

References


