Abstract

We consider a model to allocate stock levels at warehouses in a service parts logistics network. The network is a two-echelon distribution system with one central warehouse with infinite capacity and a number of local warehouses, each facing Poisson demands from geographically dispersed customers. Each local warehouse uses a potentially different base stock policy. The warehouses are collectively required to satisfy time-based service targets: Certain percentages of overall demand need to be satisfied from facilities within specified time windows. These service levels not only depend on the distance between customers and the warehouses, but also depend on the part availabilities at the warehouses. Moreover, the warehouses share their inventory as a
way to increase achieved service levels, i.e. when a local warehouse is out of stock, demand is satisfied with an emergency shipment from another close-by warehouse. Observing that the problem of finding minimum-cost stock levels is an integer nonlinear program, we develop an implicit enumeration-based method which adapts an existing inventory sharing model from the literature, and makes use of a lower bound. The results show that the proposed inventory sharing strategy results in considerable cost reduction when compared to the no-sharing case and the method is quite efficient for the considered test problems.

1. Introduction

After sales customer service is one of the factors critical to the success of any company and service parts logistics is a critical component of providing quality service, especially for high technology hardware manufacturers. Service parts logistics can be defined as the management of processes, activities and spare parts in support of repairing and maintaining products. In after sales service for industries like heavily automated production systems, information systems, automobile and aircraft manufacturing, service parts are required to repair the failed systems. Downtime of critical equipment due to part failures may have serious consequences so obtaining replacement parts quickly is essential.

In today’s market, customer service has become an important dimension of competition along with price and quality. A classical measure of customer service has been fill rate, which by definition is the percentage of the random demand satisfied with the stock on hand (Hopp and Spearman, 2001, Zipkin 2000). But in the realm of service parts logistics, customers are not concerned about whether the service provider has the part in inventory or not. They mainly care
about the time taken by the service provider to provide the service. Therefore, in order to retain
the company’s current customers and to acquire new customers, prompt service is always
recommended, for which the first requirement is to have service parts readily available.
Considering the high costs of service parts, it is not economical to have a large stock of parts.
The company therefore faces a problem of determining optimal stocking levels of service parts
that minimizes its cost while providing the demanded parts to the customers within a specified
time frame. Because customers are usually scattered over a large geographical area, many
companies use an extensive distribution network of inventory locations to ensure a high service
level.

Service parts are often supplied via a multi-echelon distribution network, i.e. a hierarchical
network of stocking locations. In a multi-echelon inventory system, warehouses at higher
echelons supply parts to lower echelon warehouses. A reason to have a multi-echelon
distribution network is the need to achieve quick response time while achieving stock
centralization to reduce holding costs. In this paper, we consider a two-echelon distribution
system with one central warehouse and a number of local warehouses, under the assumption that
customer request arrive at a local warehouse according to a Poisson process. Local warehouses
are supplied with the stock from the central warehouse, which is assumed to have infinite
capacity. As most of the service parts have random and low demand, the network is assumed to
be controlled by base stock policy in which each local warehouse continuously monitors its
corresponding inventory level. Whenever the inventory level falls below the base stock level due
to demand, the warehouse orders the service part from the central warehouse one at a time. The
main reason for using the base stock policy is the high price and low demand characteristics of the service parts.

This paper’s method seeks to determine the optimal stocking policy (optimal base stock levels at all local warehouses) that meets the service level constraints at minimal costs. Customer focus on time-based service has escalated the need of tiered service levels that are to be achieved within certain time windows (e.g., 60% of the demand satisfied within 4 hours, 80% within 8 hours). Hence, service levels not only consider part availability (fill rate) but also the response time which is a function of distance between the customer requesting the part and the warehouse that provides that part. When time-based service levels are used to measure customer service, satisfying the demand in a specified time window is more important than whether a specific warehouse satisfies it. To satisfy these stringent time-based service levels, the traditional hierarchical pass up structure (i.e., passing up the demand that cannot be satisfied by a local warehouse to the central warehouse in emergency basis) has been replaced by the network in which all the locations in the same echelon share their inventory. This leads to the concept of emergency lateral transshipments (ELTs) and direct deliveries.

ELT is passing demand along to a neighboring warehouse when a customer’s local warehouse (usually the closest to customer) does not have the part in stock. The neighboring local warehouse in the same echelon can provide the part and contribute to service. Whenever a local warehouse faces a customer demand, the demand is satisfied from the stock on hand. If the local warehouse is out of stock, the demand is satisfied from an ELT at any other local warehouses that are close. If the required stock is unavailable at all the facilities, then an emergency shipment
(direct delivery) is ordered from the central warehouse. ELTs make each stocked item more valuable as it can be used to satisfy demand not only from the service region of the warehouse where it is stocked, but also from the service region of a neighboring warehouse. The inclusion of inventory sharing makes the stocking level decisions more challenging and increases the cost of analysis. But this is compensated by the improvement in time-based service levels and by the reduced system-wide stock level necessary to obtain a certain service level. To analyze the system, we use a Markov chain model in which all the local warehouses behave as one aggregate warehouse, which is standard in the ELT models from the literature (see, e.g., Alfredsson and Verrijdt, 1999). The model is used to calculate the fraction of demand satisfied from the stock on hand and the fraction of demand satisfied through ELT for each warehouse, and the fraction of demand satisfied through direct delivery from the central warehouse.

The problem then can be stated as follows: Given a two-echelon distribution system with one central warehouse with infinite capacity and a number of local warehouses facing Poisson demands, find a set of optimal stocking levels at the local warehouses that meet all the time-based service level constraints at minimum total cost (inventory holding, transportation and penalty). The key steps to solve the problem are:

- We formulate a model of a two-echelon system in which inventory flows from a central warehouse to \( n \) local warehouses. All the local warehouses share their inventories in order to satisfy the customer demand and time-based service levels are used as a measure of customer service.

- We develop an algorithm that enumerates all the possible stocking policies and estimates the total cost for the policies which satisfy the time-based service level constraints. To
make the algorithm more efficient, we develop a lower bound so that explicit enumeration of all the possible stocking policies is not required.

• We analyze the sensitivity of the total cost with respect to each of the influencing parameters. We further investigate the advantage of obtaining lower bounds.

The paper is organized as follows: Section 2 contains the literature review on multi-echelon inventory models for service parts that consider emergency lateral transshipments. In Section 3, we formulate the problem and give the details of our modeling technique. Section 4 explains the methodology developed to solve the problem. Section 5 explains the experimental design used to test the model. We discuss the computational results in Section 6. In Section 7, we give a few concluding remarks and list some possible extensions of the current work.

2. Literature Review

Considering the vast literature on inventory management, we focus our review on inventory models for service parts that consider emergency lateral transshipments. Almost all papers in this area assume Poisson demand and one-for-one replenishment policy (also known as base stock policy or \((S-1, S)\) replenishment policy, where \(S\) denotes the base stock level). Considering the high costs and low demands of these parts in real settings, both assumptions are appropriate and representative of the practice in service parts logistics.

Considering one-for-one replenishment policies, Sherbrooke (1968) was first to analyze a multi-echelon inventory model for repairable items. He developed an approximate technique (METRIC) to calculate the stock levels at each echelon of the multi-echelon network. In the
event of a stock-out, his model assumed that the demand is backordered. Muckstadt and Thomas (1980) showed the importance of using multi-echelon inventory systems as opposed to single-echelon methods. They took data from a large industrial inventory system and used it to provide comparative results on overall inventory required to meet a given service level. They found that multi-echelon setting requires much lower total investment than the single-echelon setting achieving the same service levels and concluded that for low demand items the importance of multi-echelon approach is substantial and the importance increases as the number of low demand items becomes large. Graves (1985) developed a multi-echelon inventory model for a repairable item, which assumes that the failures occur according to a compound Poisson process and the shipment time between warehouses is deterministic. He presented a general framework for determining the steady-state distribution of the net inventory level at each warehouse. Cohen et al. (1986, 1989, 1992) propose algorithms for similar multi-echelon low-demand service parts systems.

Cohen et al. (1999) performed a benchmark study of after-sales service logistics systems for technologically complex high value products. They provided a detailed introduction into service parts logistics and reported some of the latest industrial practices and trends in service logistics operations. Two successful applications in real SPL systems are discussed in Cohen et al. (1990) and Cohen et al. (1999).

Note that in our model stock levels (hence fill rates) are decision variables which are varied to achieve the overall time-based service level. Hence, the defined (time-based) service level is a function of fill rates which were traditional “service” measures in the literature. In this sense, the
time-based service levels are similar in spirit and form to the demand-weighted composite fill rates in Muckstadt (2005) and Hopp, Spearman, and Zhang (1997). In a recent text on service parts logistics, Muckstadt (2005) collected his and others’ research results in a detailed manner. As part of the development in the book, he introduced models for allocation of inventory in a multi-echelon, multi-item distribution network and considered inventory sharing between facilities. He also considered time-based service level commitments in a different manner than ours, interpreting them as wait times due to backorders. He also described several algorithms to determine near-optimal inventory levels that meet all service level constraints at minimum investment.

Lee (1987) extended the existing low-demand models by allowing emergency lateral transshipments between facilities. He considered a two-echelon model in which the facilities were supplied from a central warehouse that was in turn supplied from a plant that was assumed to have infinite supply. He developed a procedure to approximately calculate the portion of demand met by stock on hand and the portion of demand met by emergency transshipments. He also gave an algorithm for determining the optimal stock levels at each warehouse, such that costs are minimized subject to the service level constraints. In his multi-echelon inventory model, Axsater (1990) assumed Poisson demand and one-for-one replenishments. In the event of a stock-out at a local warehouse, the replenishment is obtained from a nearby local warehouse, instead of the central warehouse. Though the system was similar to that of Lee (1986), his modeling approach was different. The demand at a local warehouse was calculated as the sum of the regular demand and the demand due to emergency lateral transshipments. He compared the results of his iterative approach with Lee’s earlier results and found that his technique was more
effective. Sherbrooke (1992a) allows facilities to have backordered lateral transshipments, which are not allowed in the previous models, focusing on simulation as a modeling technique. Most of these models assume zero or negligible lateral transshipment times. Two exceptions are Wong et al. (2005a) who consider nonzero lateral transshipment times, and Tagaras and Dimitrios (2002) who investigate the effects of demand distributions along with nonnegligible transshipment times. Another variation of these models is Wong et al. (2005b) that considers downtime cost explicitly, but ignores service level constraints.

The model presented in this report is an extension of the model analyzed by Alfredsson and Verrijdt (1999), which builds on Axsater (1990). They considered a two-echelon inventory system where the demands at facilities were not backordered, but satisfied through an emergency lateral transshipment, a direct delivery from a central warehouse, or a direct delivery from a manufacturing facility. They developed an iterative procedure to calculate the fill-rates at each local warehouse, which incorporates the probabilities of using different facilities as the ELT source for a particular local warehouse. They concluded that using both ELT flexibility and direct delivery flexibility results in significant cost reductions compared with using no supply flexibility at all. They also observed that the distribution of shipment times (assumed to be exponential) had a negligible effect on long term service performance. In the model, the central warehouse had a limited supply and received replenishments from a manufacturing facility that was assumed to have infinite production capacity. Their model did not illustrate a method to calculate the optimum stock levels, which we do in the current paper, in the form of implicit enumeration. Finally, we minimize the total cost subject to time-based service level constraints, which were not present in any of the earlier models.
3. Model Formulation

3.1. Description

The model presented in this report is based on the model developed by Alfredsson and Verrijdt (1999). The model considers a single item, two-echelon distribution network model for service parts, consisting of a central warehouse and a number of local warehouses. All the customers are dispersed among the local warehouses based on their relative positions, i.e., a customer would be assigned to the closest local warehouse. We assume customer demand follows an independent Poisson process, requesting one part at a time. The local warehouses are supplied with the stock from the central warehouse, which is assumed to have infinite capacity. Whenever a customer request arrives at a local warehouse, the demand would be satisfied from the stock on hand, if available. All the local warehouses follow a base stock replenishment policy in which inventory is replenished by the central warehouse one unit at a time as random demand occurs at the local warehouses.

In case of a stock-out at a local warehouse, customer demand is satisfied by an emergency lateral transshipment (ELT) from a neighboring local warehouse holding the required stock. If none of the local warehouses are able to satisfy the demand, the demand is then satisfied by a direct delivery from the central warehouse. For example, consider the case of three local warehouses (numbered 1, 2, and 3). Warehouses 2 and 3 are the neighboring locations for warehouse 1. If a customer arrives at warehouse 1 and if it is out of stock, the demand would be satisfied by an ELT from warehouse 2 or 3, if either has the required stock. If neither has the stock, the demand would be satisfied by a direct delivery from the central warehouse. We assume that there is no lost demand and the customer who initiates the ELT or direct delivery waits for the part to arrive.
even though it is possible the customer’s assigned warehouse may receive the item through normal replenishment while the customer is waiting). Figure 1 illustrates a system with one central warehouse and three local warehouses. The arrows show part supply and replenishment directions. The dotted lines show ELTs and direct deliveries.

Inventory sharing between the local warehouses make the stock more valuable as the stock is used not only for the assigned customer, but also by another local warehouse to satisfy their respective customers. This would decrease the total inventory required and hence the holding costs for a particular service level. But ELTs and direct deliveries would have higher transportation costs and there would be penalty costs associated with them, which would increase the costs. Hence, the inventory allocation becomes a more difficult problem.

Figure 1 A two-echelon system with three local warehouses and inventory sharing

Alfredsson and Verrijdt (1999) showed that their model was not sensitive to the choice of lead time distribution. To facilitate the model development, we assume that the replenishment lead
times are identically and independently distributed exponential random variables. By definition, fill rate is the probability of filling the demand with the stock on hand. Because fill rate does not capture the time taken to satisfy the demand, we define the time-based service level as the percentage of demand satisfied within a certain time window. For example, we have three time windows, 4hr, 8hr and 12hr. Hence, time-based service level depends not only on the part availability but also on the distances between warehouses and customers (whether a warehouse and its customer is within the time window or not).

Following is the summary of the notation used in the model:

- \( n \) Number of local warehouses
- \( i \) Local warehouse referred as main location, \( i = 1, 2, \ldots, n \)
- \( q \) Local warehouse referred as neighboring warehouse, \( q = 1, 2 \ldots n \) and \( q \neq i \)
- \( k \) Index for customer \( k \)
- \( l(k) \) Local warehouse assigned to customer \( k \)
- \( S_i \) Base level stock at local warehouse \( i \)
- \( S \) Total stock level (i.e., \( S = \sum_{i=1}^{n} S_i \) )
- \( \lambda_{ik} \) Demand of customer \( k \) assigned to local warehouse \( i \)
- \( \lambda_i \) Total demand rate at local warehouse \( i \)
- \( \lambda \) Total demand rate
- \( L \) Average replenishment lead time (assumed same for all warehouses)
- \( \mu \) Rate at which replenishment orders arrive at a local warehouse (\( =1/L \))
- \( e_i \) ELT demand at local warehouse \( i \) from other warehouses
- \( g_i \) Total demand at local warehouse \( i \) (after considering ELT demand)
\[ \beta_i \quad \text{Fraction of demand satisfied from on hand stock at local warehouse } i \]
\[ \alpha_i \quad \text{Fraction of demand satisfied through ELTs at local warehouse } i \]
\[ \gamma \quad \text{Fraction of demand satisfied from the central warehouse through direct deliveries} \]
\[ \omega_{iq} \quad \text{Probability of using local warehouse } q \text{ as a source for an ELT for a customer assigned to local warehouse } i \]
\[ \pi_j \quad \text{Steady state probability that there is a total of } j \text{ items on hand at all the local warehouses} \]
\[ \pi^i_j \quad \text{Steady state probability that there are } j \text{ items on hand at local warehouse } i \]
\[ c \quad \text{Unit price of the service part} \]
\[ t \quad \text{Unit cost of transporting the part from central warehouse to local warehouses} \]
\[ u_{ik} \quad \text{Unit cost of transporting the part from local warehouse } i \text{ to customer } k \]
\[ v \quad \text{Unit cost of transporting the part from the central warehouse to a customer through direct delivery (includes both transportation and penalty cost)} \]

3.2. Aggregate Model for $\gamma$

For the fraction of demand satisfied by the central warehouse ($\gamma$) we consider an aggregate model, i.e., all the local warehouses as one single warehouse with a total stock level of $S = \sum_{i=1}^{n} S_i$, and experiencing a total demand of $\lambda = \sum_{i=1}^{n} \lambda_i$. The value of $\gamma$ is the same for all the local warehouses, independent of their individual stock levels and demand rates (for a given total stock level and a given total demand). The steady state probabilities are calculated by solving the global balance equations for the CTMC (Kulkarni, 1995). Then, for $j = 1, 2, \ldots, S$, we obtain
The value of $\gamma$ is then equal to the probability of having zero total stock level, since all the local warehouses share their inventories and try to satisfy demands from their on-hand inventories before requesting for a direct delivery from the central warehouse. Thus,

$$\gamma = \frac{1}{\sum_{j=0}^{S} \left( \frac{S!}{(S-j)!} \left( \frac{\mu}{\lambda} \right)^j \right)}.$$  

### 3.3 Model for ELTs and Local Fill Rates ($\alpha_i$ and $\beta_i$)

As described by Axsater (1990), the state space at each local warehouse can be modeled as a CTMC. In this model, the demand rate at each local warehouse is adjusted by taking into account the demand coming from other local warehouses in the form of ELTs. For a local warehouse facing a regular demand $\lambda_i$ and an ELT demand of $e_i$, the adjusted demand rate can be calculated as $g_i = \lambda_i + e_i$. The demand process at local warehouse $i$ is still assumed to be Poisson with rate $g_i$. Thus, the analytical solution to the steady state equations of this model can be obtained similarly. The following solution is obtained for steady-state probabilities:

$$\pi^{i}_j = \left( \frac{S!}{(S-j)!} \right) \left( \frac{\mu}{g_i} \right)^j \pi^{i}_0$$ for $j = 1, 2 \ldots S$, 

where

$$\pi^{i}_0 = \frac{1}{\sum_{j=0}^{S} \left( \frac{S!}{(S-j)!} \left( \frac{\mu}{\lambda_i} \right)^j \right)}.$$
where \( \pi^i_0 = \frac{1}{\sum_{j=0}^{S_i} \left( \frac{S_i!}{(S_i-j)!} \right) \left( \frac{\mu}{g_i} \right)^j} \)

As \( \pi^i_0 \) denotes the probability of not having a stock at warehouse \( i \), we can compute the fill rates easily: \( \beta_i = 1 - \pi^i_0 \), and \( \alpha_i = 1 - \beta_i - \gamma \).

ELT demand rate \( e_i \) still needs to be evaluated and it is done following the approach described by Alfredsson and Verrijdt (1999). Let \( f_i \) denote the fraction of total demand \( \lambda \) that is satisfied at local warehouse \( i \). Then \( f_i \lambda \) is equal to \( \lambda + e_i \) if warehouse \( i \) has stock on-hand and \( f_i \lambda \) is equal to zero if warehouse \( i \) is out of stock. Therefore, \( f_i \lambda = \beta_i (\lambda + e_i) + (1 - \beta_i) 0 \). Fraction \( f_i \) can also be interpreted as the sum of the fraction \( \beta_i \) of the demand \( \lambda_i \) originating at local warehouse \( i \) and the fraction of ELT demand of all other local warehouses that is satisfied by local warehouse \( i \). Then

\[
f_i \lambda = \beta_i \lambda_i + \sum_{q=1}^{n} q^* q_i^* q_\lambda \lambda_q .
\]

From above two equations, we obtain the following expression for \( e_i \):

\[
e_i = \frac{1}{\beta_i} \sum_{q=1}^{n} q^* \alpha_i q_\lambda \lambda_q .
\]

Clearly, \( e_i \) is a function of both \( \alpha_i \) and \( \beta_i \). Hence, we use an iterative procedure to evaluate \( \alpha_i \) and \( \beta_i \), where we alternatively update the values for \( e_i \), \( \alpha_i \) and \( \beta_i \). Axsater (1990) showed that the convergence is obtained in a few iterations. Iterations are started by setting \( e_i \) to zero for all \( i \) and at each iteration the \( e_i \)'s are used to calculate the adjusted demand rates \( g_i \), which are further used to calculate the steady state probabilities at each warehouse. Then, \( \alpha_i \)'s and \( \beta_i \)'s are calculated.
and are used to recalculate $e_i$’s. New values of $e_i$’s are used in the next iteration and the whole process is continued until convergence. For local warehouse $i$ having a stock level of zero, we know that $\beta_i$ is zero and hence $\alpha_i = 1 - \gamma$. So, the iterative procedure is not required for warehouses with a stock level of zero.

We now need to provide a procedure to evaluate $\omega$’s. Note that for $n = 2$, this probability is always equal to 1. Also, when a randomly chosen neighbor with stock on-hand is used to source an ELT, the probability $\omega_{iq}$ is equal to $1/(n - 1)$ for all $i$ and $q$.

### 3.4 Evaluation of Time-based Service Levels

Time-based service levels are defined as the percentage of total demand satisfied within a specified time window. In our model, only three time windows (e.g., 4 hr, 8 hr and 12 hr) are considered. The central warehouse is assumed to be located far away from the customers and hence the direct deliveries from the central warehouse do not affect any of the time-based service levels under consideration.

For each local warehouse, all the customers (including the customers in the service region of other warehouses) are categorized such that the customers who are within 4 hr from the warehouse make up the 4 hr customers, customers who are within 8 hr from the warehouse make up the 8 hr customers, and so on. The time taken by a local warehouse to provide service to a customer depends on the distance between the local warehouse and the customer.

We introduce the following additional notation to calculate the time-based service levels:

$W_{ik}$: The time window in which customer $k$ can be satisfied by a delivery from facility $i$. 
\( \Lambda_{wi} \): Total demand that is in the time window \( w \) at local warehouse \( i \), i.e. \( \sum_{k \in \{k\mid k \neq w_i\}} \lambda_{ik} \).

\( \Lambda'_{wq} \): Total demand of all customers of local warehouse \( i \) for which, warehouse \( q \) \((q \neq i)\) is in the time window \( w \), i.e. \( \sum_{k \in \{k\mid k \neq w_i\}} \lambda_{ik} \).

\( F_w \): Time-based service level for time window \( w \).

The time-based service level for a particular time window \( w \) can be calculated as the ratio of the total demand that is satisfied in time window \( w \) to the total demand \( \lambda \). The demand that is satisfied in time window \( w \) is divided into two parts: The demand satisfied by on-hand inventory and demand satisfied by ELTs. Assuming local warehouses satisfy demands in first come, first served fashion (i.e., without any prioritization of customers), demand that is satisfied in time window \( w \) by on-hand inventory at local warehouse \( i \) can be determined by \( \beta_i \Lambda_{wi} \).

Recall \( \omega_{iq} \) is the probability of using local warehouse \( q \) as a source for an ELT for a customer assigned to warehouse \( i \) and \( \Lambda'_{wq} \) represents the total demand of all customers of warehouse \( i \) for which warehouse \( q \) \((q \neq i)\) is in time window \( w \). Hence, the amount \( \omega_{iq} \Lambda'_{wq} \) gives the total demand of local warehouse \( i \) that can be satisfied by a neighboring warehouse \( q \). Therefore, we calculate \( \sum_{q \neq i} \omega_{iq} \Lambda'_{wq} \) which yields the total demand at local warehouse \( i \) that can be satisfied in
time window $w$ by all the neighboring local warehouses. Multiplying $\sum_{q+1 \neq i = 1}^{q} \theta_{iq} \Lambda_{wiq}^i$ by $a_i$, we obtain the total demand of warehouse $i$ that is satisfied in time window $w$ by ELTs (again assuming FCFS demand satisfaction). Thus, we obtain the following expression for $F_w$:

$$F_w = \frac{1}{\lambda} \sum_{i=1}^{n} \left( \beta_i \Lambda_{wi}^i + a_i \sum_{q+1 \neq i = 1}^{q} \theta_{iq} \Lambda_{wiq}^i \right), \quad \text{for } w = 4, 8 \text{ and } 12.$$ 

Note that time-based service level is a system-wide measure that includes all the warehouses and hence is a function of stock levels and customer demands of all warehouses. It further depends on the distances of all the customers with respect to all local warehouses. In the computational study, base-level time-based service targets form the constraints as 60% of the total demand should be satisfied in 4 hours, 80% in 8 hours, and 95% in 12 hours.

### 3.5. Total Cost Function

The cost of the system depends on the observed fill rates of the local warehouses for a given stock distribution. The total cost function includes several components:

**Inventory Holding Cost:** We measure the holding cost as the holding cost per unit item per unit time ($h$) times the average inventory across all local warehouses. Using the steady state probabilities for the system (recall that $\pi_s$ is the probability of having a total of $s$ items on hand at all the local warehouses), we can calculate the average inventory across all local warehouses using the lost-sales average inventory formula for a virtual facility with a total stock level $s$ as follows:
Average Inventory = $S - (1 - \gamma) \frac{\lambda}{\mu}$

We can do this as all the available inventory in the system is fully shared by all local warehouses. Then, the inventory holding cost per unit time for the overall system is

$$h\left(S - (1 - \gamma) \frac{\lambda}{\mu}\right).$$

**Transportation and Penalty Costs:** Transportation cost includes the cost of transporting the part from the central warehouse to local warehouses and from local warehouses to the customers. In the system described here, a customer ends up being served by its assigned warehouse, a neighboring local warehouse, or the central warehouse. Hence, multiple options for supplying parts to customers need to be considered. Whenever a customer is satisfied from an ELT or a direct delivery from the central warehouse, then there is some loss of customer goodwill which is captured by a penalty cost. In the model, penalty costs have been in transportation costs only, by multiplying the transportation costs by a factor.

Transportation costs are further divided into four categories:

i. **Cost of transporting the part from the central warehouse to local warehouses** ($TC_1$): The central warehouse is assumed to be far-off from all the local warehouses and customers, and therefore, the cost of transporting the part from the central warehouse to any of the local warehouses is considered to be same. Therefore, $TC_1 = t(1 - \gamma)\lambda$.

ii. **Cost of transporting the part from local warehouse $i$ to customer $k$ when $l(k)=i$** ($TC_2$): For a local warehouse $i$, $\beta_i$ represents the fraction of demand satisfied from on-hand stock.
Hence, \( \sum_k \beta \lambda_{ik} \) is the total demand warehouse \( i \) satisfies directly and the cost of satisfying that demand is \( \sum_k \beta \lambda_{ik} u_{ik} \). Therefore, \( TC_2 = \sum_i \sum_k \beta \lambda_{ik} u_{ik} \).

iii. \textit{Cost of transporting the part from local warehouse} \( q \) to customer \( k \) when \( l(k) \neq q \) (\( TC_3 \)): This is the cost of emergency lateral transshipments. To include penalty costs, the transportation cost of ELTs is multiplied by a factor \( \rho \), which is assumed to be independent of \( q \) and \( k \). This implies that \((\rho - 1)\) times the transportation cost of ELTs is the penalty cost incurred because of ELTs. Recalling \( \alpha_i \) denotes the fraction of demand satisfied through ELTs at local warehouse \( i \) and \( \omega_{iq} \) is the probability of using local warehouse \( q \) as a source for an ELT for a customer assigned to local warehouse \( i \). Hence,

\[ \sum_k \omega_{iq} \alpha_i \lambda_{ik} \] represents the total demand of warehouse \( i \) satisfied by warehouse \( q \) and the cost of satisfying that demand is \( \sum_k \omega_{iq} \alpha_i \lambda_{ik} u_{ik} \). Therefore, \( TC_3 = \rho \sum_i \sum_{q,q'} \sum_k \omega_{iq} \alpha_i \lambda_{ik} u_{ik} \).

iv. \textit{Cost of direct deliveries from the central warehouse to customer} \( k \) (\( TC_4 \)): Since customers are assumed to be far-off from the central warehouse, both the transportation and penalty costs are independent of the customer and are considered in the single cost parameter \( v \). The product \( \gamma \lambda \) is the total demand satisfied through direct deliveries from the central warehouse. Therefore, \( TC_4 = v \gamma \lambda \)

Hence the total cost can be evaluated as follows:
Total cost = \sum_i \left( S_i - \frac{\lambda_i}{\mu} \right) + \kappa (1-\gamma) \lambda + \sum_i \sum_k \beta_i \lambda_{ik} u_{ik} + \rho \sum_i \sum_{q,q'} \sum_k \omega_{iq} \lambda_{ik} u_{qk} + \nu \gamma \lambda

To achieve high service levels, high stock levels are required, which would result in high holding costs and both \( TC_1 \) and \( TC_2 \) would increase as more demand would be satisfied by on hand inventory and a less number of direct deliveries and ELTs would be required. Hence, both \( TC_3 \) and \( TC_4 \) would decrease. Therefore, an optimization method is needed to find the optimal stock levels that minimizes the total cost and at the same time, satisfies the time-based service level constraints. The cost minimization problem in the steady state is:

\[
\text{Min} \ Z = \sum_i \left( S_i - \frac{\lambda_i}{\mu} \right) + \kappa (1-\gamma) \lambda + \sum_i \sum_k \beta_i \lambda_{ik} u_{ik} + \rho \sum_i \sum_{q,q'} \sum_k \omega_{iq} \lambda_{ik} u_{qk} + \nu \gamma \lambda
\]

subject to

\[
F_w = \frac{1}{\lambda} \sum_{i=1}^q \left( \beta_i \Lambda_{wi} + a \sum_{q'=1}^q \omega_{iq} \Lambda^{q'}_{wiq} \right) \geq F_w w = 4, 8, 12,
\]

\( S \geq 0 \), and integer, \( \forall i \)

where \( F_w \) is the target time-based service level for time window \( w \). In the experiments, expressed as percentages, we use \( F_4 = 0.6 \), \( F_8 = 0.8 \), and \( F_{12} = 0.95 \), as base settings. To calculate the total cost objective function \( (Z) \) and constraints (time-based service levels), values of \( \gamma, \alpha_i, \beta_i \) (for all \( i \)) and \( \omega_{ij} \) (for all \( i, j \)) are required. Recall that \( \alpha_i, \beta_i \) and \( \omega_{ij} \) are obtained by using the iterative procedure described earlier.

3.6 No Inventory Sharing Case

For the case when there is no inventory sharing between the local warehouses, the aggregate model is not valid and all the local warehouses behave independently of each other. So the
fraction of demand that is satisfied by the central warehouse for a local warehouse \( i (\gamma_i) \) depends on the stock level and demand rate at local warehouse \( i \). Since there is no sharing \( \alpha_i \) is zero for all \( i, \beta_i \) can be calculated as \( 1 - \gamma_i \) and hence iterations to calculate the fill-rates are not required in this case. Therefore,

\[
\gamma_i = \frac{1}{\sum_{j=0}^{S_i} \left( \frac{S_i!}{(S_i - j)!} \right) \left( \frac{\mu}{\lambda_i} \right)^j}, \forall i
\]

\[
\beta_i = 1 - \gamma_i, \forall i.
\]

The total cost can be found by just adding the costs of each of the local warehouses. There is no ELT transportation cost as there is no sharing. Also, the approximation for the holding cost used in the sharing case is not used here, as it will result in a negative holding cost for the local warehouse having zero stock level. The equations used for the no-sharing case are as follows:

\[
\text{Total Cost} = H \left[ \sum_{i=1}^{n} \left( \sum_{s=0}^{S_i} \frac{S_i!}{(S_i - s)!} \left( \frac{\mu}{\lambda_i} \right)^s \right) + \left( t(1 - \gamma_i) \right) \right] + \left( \sum_{k} \beta_i \lambda_i u_{ik} \right) + \left( \gamma_i \lambda_i \right)
\]

\[
F_w = \frac{1}{k} \sum_{i=1}^{n} \beta_i \lambda_i w_{i}, \text{ for } w = 4, 8, 12.
\]

4. Solution Methodology

The optimization problem described in the previous chapter is non-linear with implicit constraints and procedures and is quite complicated to solve. If the stock level at each local warehouse is known, it is easy to evaluate the total cost and to check whether time-based service
level constraints are satisfied. Hence, we develop a method based on enumeration, in which all possible stock profiles (stock levels across all local warehouses, \((S_1,\ldots,S_n)\)) are enumerated and the total cost for each enumeration is calculated. For each stock profile, we check whether it satisfies the time-based service level constraints or not. The optimal stock profile is the one that satisfies all of the time-based service level constraints and yields minimum total cost. The algorithm is coded in C++ and is run on Pentium M 1.4 GHz Centrino processor with a Windows-based operating system.

Because service parts have very low demand, we assume that in the optimal solution a local warehouse would never stock more than three parts. Therefore, each local warehouse can have a stock level of zero, one, two or three. Hence, for a 10-warehouse problem, approximately a million \(4^{10}\) stock profiles (combinations of individual stock levels across all warehouses) are possible and the objective is to find the one that is optimal. Considering that enumerating a million configurations and then calculating their costs (and service levels), is not particularly efficient, we introduce a lower bound that tries to eliminate the unpromising stock profiles from consideration, making the overall procedure an implicit enumeration algorithm.

For a 10-local warehouse problem, the total stock across all the local warehouses can range from 0 to 30. To find the optimal stock profile, all the possible stock profiles need to be (implicitly or explicitly) enumerated for each total stock level. In order to decrease the number of enumerations required, a lower bound on cost is calculated corresponding to every total stock level. The lower bound is calculated by setting \(\alpha_i = 0\), implying that \(\beta_i = 1 - \gamma\) for all \(i\) (i.e., \(\gamma\) is calculated assuming full sharing among local warehouses, and all ELT sharing \((\alpha_i)\) is assumed
part of local fill rates ($\beta_i$)). Therefore, the lower bound on cost for each total stock level ($S$) is calculated as

$$LB = h \left( S - \frac{\lambda}{\mu} \right) + t(1 - \gamma)\lambda + \beta \sum_i \sum_k \lambda_i u_{ik} + v \gamma \lambda,$$

where

$$\gamma = \frac{1}{\sum_{j=0}^{S} \frac{S!}{(S-j)!} \left( \frac{\mu}{\lambda} \right)^j}, \text{ and } \beta = 1 - \gamma.$$

Lower bounds for all the total stock levels (say 0 to 30) are then sorted. We start by enumerating all the stock profiles for the total stock level having the lowest lower bound and then for the one having second lowest lower bound and so on. We do this hoping that enumerating stock levels with smaller lower bounds first would give good feasible solutions that can be used to eliminate nonpromising stock profiles that have larger lower bounds than the best feasible solution’s cost. For each enumerated stock profile, $a_i$ and $\beta_i$ are calculated (for all $i$) using the iterative procedure. Then, the total cost is calculated and the time-based service level constraints are tested. Whenever a feasible stock profile is enumerated, if it is better than the best solution found so far, then its total cost is set as the new incumbent upper bound on the (optimal) total cost. We obtain an upper bound on cost (best solution achieved so far) after enumerating all the stock profiles for a particular total stock level.

At the start of enumeration for a total stock level if it is found that the current upper bound is less than the lower bound for the total stock level, then there is no need of further (stock profile) enumeration of the total stock level. All the remaining total stock levels would also have lower
bound greater than the upper bound as they are already sorted. The stock profile corresponding to the most recent upper bound provides the optimal stock levels at all local warehouses and the upper bound is reported as the total cost corresponding to the optimal solution.

5. Experimental Settings

To investigate the behavior of the system and the performance of the algorithm with changes in the system parameters, a hypothetical network is designed by randomly locating 75 customers and 10 local warehouses on a grid of $100 \times 100$ ‘miles’. Four different networks are designed so that the results are not biased by the location of customers and local warehouses. We study the effect of variation in the following system parameters:

- Customer demands
- Replenishment rate (or lead time)
- Holding cost per unit item
- Penalty factor ($\rho$)
- Unit cost of direct delivery from central warehouse ($v$)
- Unit cost of delivery of part from central warehouse to a local warehouse ($t$)

Each customer demand is assumed to follow a Poisson distribution. To study the effect of change in demand rates, first a number between 0 and 0.15 is generated for each customer. These are considered to be the means of the demand rate for the customers. This gives one set of demand rates. By multiplying and dividing these demand rates by 1.5, two other sets of demand rates are generated. The demand is chosen in this range so as to incorporate low-demand characteristic of service parts logistics systems. The lead times are assumed to be exponentially distributed. Three
different values for the mean of the exponential function are considered: 2 days (replenishment rate of 15 times per month), 3 days (replenishment rate of 10 times per month), and 6 days (replenishment rate of 5 times per month).

For each of the 75 customers and 10 local warehouses, the $x$ and $y$ coordinates are determined by generating two random numbers between 0 and 100. Using the $x$ and $y$ coordinates, the Euclidean distance of the customers from every local warehouse is calculated. Each customer is then assigned to its nearest local warehouse. Assuming that cost of transporting a part from each local warehouse to every other customer ($u_{ik}$) is directly proportional to distance between them, the cost of transportation is assumed to be $2.5 per unit distance.

To incorporate the penalty costs associated with ELTs, the transportation cost of ELTs is multiplied by factor $\rho$, which is varied from two to four. The cost of direct deliveries from the central warehouse is assumed to be independent of customer location and is generally very high compared to normal shipment costs. Incorporating the penalty costs as well, the unit cost of direct deliveries from the central warehouse is varied between $500 and $1500. The effect of cost of transporting the part from central warehouse to a local warehouse ($t$) is studied by varying it between $10 and $30. The cost of the part and interest rate are required to calculate holding costs. As service parts are generally expensive, the cost of the part is varied between $500 and $4500. The monthly interest rate has been fixed at 2%.

The time window in which the customer can be satisfied will depend on the distance between the customer and local warehouse. Therefore, for each local warehouse, all the customers are
grouped into four categories: customers who can be satisfied within 4 hr, 8 hr, 12 hr and more than 12 hr from the warehouse. It is assumed that service can be provided in 4 hrs if the distance is less than 20 ‘miles,’ in 8 hrs if distance is less than 40 miles, and in 12 hrs if distance is less than 60 miles.

In the model, time-based service levels form the constraints because 60% of the total demand should be satisfied in 4 hours, 80% in 8 hours, and 95% in 12 hours.

6. Computational Results

More than 200 individual problems are solved using the algorithm introduced in Section 4. The implicit enumeration algorithm solves all instances within 3 minutes of CPU time. In all the problem instances the optimal stock level at a local warehouse was always less than or equal to 2 parts. This shows that having more than a stock of 3 parts at a local warehouse as an option would never be optimal for the generated demands. The optimal stock levels vary across different networks for the same values of parameters mainly due to time based service levels which depend on relative locations of customers and warehouses. Hence, we conclude that the locations of local warehouses and customers can have an impact on the stocking decisions. We present a summary of the results here.

6.1 Effect of the lower bounds

As discussed earlier, there are more than a million \(4^{10}\) enumerations possible for a 10-local warehouse problem. Our results show that the number of enumerations required to find the optimal stock profile within the algorithm is reduced significantly. For our model, the total
number of enumerations required is reduced by at least 65%, in some instances much more, by up to 94%. These numbers justify the use of the lower bounds. The number of enumerations required increases as the unit holding cost \( (h) \) decreases. This is because the lower bounds are calculated by neglecting the ELT costs and hence, the lower bounds would not be as tight if the holding costs are very low as compared to cost of the ELTs.

6.2 Effects of the direct delivery cost \((v)\)
If there is significant sharing between local warehouses, the fraction of demand that is satisfied by the central warehouse \((\gamma)\) would be low. Hence, total cost of direct deliveries \((TC_4)\) would have negligible effect on the cost of the system or on the amount of total stock required. This is a general trend for all the problem instances. Therefore, the cost of direct delivery is considerable only when the total stock level is low. However, low (total) stock levels will not usually be feasible because of the time-based service level constraints. Hence, there is no need to consider direct delivery costs explicitly in the model if there are stringent time-based service level constraints. On the other hand, for the no-sharing case, the fraction of demand satisfied by the central warehouse can be significant and hence, there is a substantial effect of the direct delivery cost on the system, both its cost and its service.

6.3 Effects of the transportation cost from the central warehouse to local warehouses \((t)\)
Increase in \(t\) increases cost of transportation from the central warehouse to the local warehouse \((TC_1)\), but has no effect on other costs. Hence, the total cost increases with increase in \(t\). Changes in \(t\) do not affect the optimal stock levels for the system. We can explain this using the formula for \(TC_1\) and \(TC_4\) derived earlier:
\[ TC_1 + TC_4 = t(1 - \gamma) \lambda + v \gamma \lambda = t \lambda + (v - t) \gamma \lambda , \]

which is approximately equal to \( t \lambda + v \gamma \lambda , \) as \( v \gg t. \) Then, we have

\[ \frac{\partial (TC_1 + TC_4)}{\partial t} = \lambda . \]

Therefore, the effect of changing \( t \) on total cost depends mainly on \( \lambda \) and stocking decisions are not affected by the value of \( t. \) The above derivation assumes that the unit direct delivery cost \( (v) \) is much larger than the cost of transportation from the central warehouse to local warehouses \( (t). \) This is generally true as the penalty cost for using the central warehouse for a direct delivery to a customer is very high.

### 6.4 Effect of the unit holding cost \( (h) \)

From Figure 4, it is clear that increase in holding cost increases the total cost of the system. It also shows that the total cost is reduced by a considerable amount by inventory sharing. Our results show that increasing the holding cost decreases the optimal total stock. Average inventory decreases because of decrease in optimal stock levels. Considering the increase in unit holding cost, the net effect is an increase in the total holding costs. A decrease in total stock also decreases the cost of transporting parts from the local warehouses to their customers \( (TC_2), \) but increases the cost due to ELTs \( (TC_3) \) and direct deliveries \( (TC_4). \) The results also show that after a certain level, an increase in holding cost does not affect the optimal total stock. This happens because of the time-based service level constraints, which force the system to have certain a minimum amount of stock.
6.5 Effects of the penalty factor (\(\rho\))

The results suggest that the total cost of the system increases with the increase in penalty factor (Figure 5). The optimal stock level increases as the penalty factor goes up. Intuitively this makes sense, as the system would try to satisfy more demand from the assigned local warehouses rather than using an ELT. An increase in the optimal stock level increases holding costs and \(TC_2\), and decreases costs due to ELTs (\(TC_3\)) and direct deliveries (\(TC_4\)).

Figure 2. Effect of changing unit holding cost

Figure 3. Effect of changing penalty factor
6.6 Effects of the replenishment rate ($\mu$)

Figure 6 depicts that the total cost decreases with an increase in the replenishment rate or a decrease in the lead time ($1/\mu$). As the lead time decreases, local warehouses are replenished faster, and hence they can have a lower base stock level, which would lead to a decrease in total cost. Similar to the case of holding costs, after a certain level, an increase in the replenishment rate does not affect the optimal total stock. This happens because the time-based service level constraints force the system to have certain minimum amount of stock. There is again a substantial decrease in the total cost, if sharing between the local warehouses is allowed.

![Sharing without Priority](image)

![No Sharing](image)

Figure 4. Effect of changes in Replenishment Rate

6.7 Effects of total demand ($\lambda$)

Figure 7 shows that the total cost of the system increases with increase in total demand. As the demand increases, the local warehouses need to have higher base stock levels and the total cost of the system increases. In almost every case, all cost types increase with an increase in demand (in turn, increase in the base stock level). Again, Figure 7 shows that sharing is much better than no-sharing.
6.8 Effects of time-based service levels

To study the effect of relaxing the time-based service level constraints, the constraints are changed as 40% of the total demand should be satisfied in 4 hours, 60% in 8 hours and 85% in 12 hours. Our results show that the optimal solution does not change if these relaxations are applied. The reason is the relative difference between the unit holding cost and the unit cost of an ELT. With the lower service levels, smaller base stock levels become feasible but the lowest cost solution does not change. When the same lower service levels are run for higher unit holding cost, the optimal solution changes. The solution suggests lower total stock level for the system and obtains a lower total cost. The decrease in total cost is more substantial in the case of sharing.

7. Conclusions and Future Work

In this paper, we consider a two-echelon service parts logistics system with one central warehouse and a number of local warehouses. The problem is to determine the optimal base stock levels at all local warehouses that meet all the time-based service level constraints at
minimal costs. We develop an implicit enumeration based algorithm to optimally solve the problem. To satisfy these stringent time-based service levels, we allow all the warehouses in the same echelon to share their inventory in the form of emergency lateral transshipments.

The results indicate that to achieve a comparable level of service without location interactions, one needs higher base stock levels, and the total system cost would be considerably higher than the case in which inventory sharing is allowed. We study the effects of changing various system parameters.

There are several possible extensions and improvements of the model. One can extend the model to include multiple parts. In such a model, each part can have its own target service level, holding cost and penalty cost depending on criticality. We also assume fixed locations for warehouses. One can integrate facility location decisions into this model. We can also extend the model to consider a variety of transportation modes that can be used to satisfy the demand. Also, we can make the model more flexible by allowing the facilities to have the option of sharing inventory in a particular scenario or not (instead of forcing inventory sharing for all facilities in all scenarios). Finally, we can extend the model by dividing the whole region into multiple service regions, each potentially with different costs and service levels.

References


Cohen, M.A., P.R. Kleindorfer, H.L. Lee, and D.F. Pyke, 1992, Multi-item service constrained (s,S) policies for spare parts logistics systems, Naval Research Logistics, 39, 561-577.


