Modeling emergency supply flexibility in a two-echelon inventory system

Modeling Emergency Supply Flexibility in a Two-Echelon Inventory System

Jos Verrijdt and Patrik Alfredsson

Research Report TUE/TM/LBS/96-01
February, 1996

MODELING EMERGENCY SUPPLY FLEXIBILITY IN A TWO-ECHELON INVENTORY SYSTEM

Patrik Alfredsson[a] and Jos Verrijdt[b]

[a] Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

[b] Faculty of Technology Management, Paviljoen F 12, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Abstract

We consider a two-echelon inventory system for service parts. To obtain high service levels at a low cost we allow not only for normal supply of parts but also for emergency supply options in terms of lateral transhipments and direct deliveries. After presenting the strategy we use for satisfying customer demand, we construct an analytical model which we use to calculate relevant performance measures. Simulation shows that our model produces accurate estimates, and that the performance of the inventory system is insensitive to the leadtime distribution. After introducing a cost structure we show that the strategy we propose can result in considerable savings when compared to using only normal supply.

Keywords: inventory control; service parts; emergency supply

1 Introduction

After-sales service has become more and more a competitive weapon in the nineties. In various kinds of industry, such as car manufacturing, information systems and communication systems, aircraft manufacturing and many others, fast and reliable supply of service parts to customers is crucial in order to retain current customers or to obtain new customers. Since these customers are usually scattered over a large geographical area, many companies use an extensive distribution network of inventory locations in order to guarantee a high service level. The investment in inventory in such networks can be very high and therefore it could be advantageous to implement flexibility. Examples of such flexibility options are the use of emergency lateral transhipments and direct deliveries. Lateral transhipments are used to fill a demand at a local warehouse that is out of stock from any other local warehouse that does have stock on hand. Direct deliveries are used to fill such a demand from a higher level in the system, e.g., a central warehouse or a plant. The trade-off that is associated with the use of these flexibility options is between costs and service performance.

In this paper we present a two-echelon model consisting of a central warehouse supplying a number of local warehouses. These local warehouses can also be interpreted as moving car stocks with which service engineers visit customer sites to replace failed service parts. The central warehouse is supplied from a plant which we assume to have infinite supply. The inventory policy applied is one-for-one replenishment. This inventory policy is very common in practice for service parts, because of the high price and low demand characteristics of many of these items. In case of a demand at a local warehouse, we apply the following strategy for filling this demand:

1. Fill the demand from stock on hand. The local warehouse where the demand occurs issues a replenishment order to the central warehouse.

2. If this is not possible, the demand is satisfied by an emergency lateral transhipment (ELT) from another, randomly chosen, local warehouse that has stock on hand. We assume that the local warehouses are situated in one geographical area such that the ELT time is much shorter than the replenishment time from the central warehouse. The local warehouse that sources the ELT issues a replenishment order to the central warehouse.

3. If this is not possible, the demand is satisfied by a direct delivery from the central warehouse if it has stock on hand. Here we assume that a direct delivery is much faster than a normal replenishment from the central warehouse. The central warehouse issues a replenishment order to the plant.

4. Finally, if this is not possible, the demand is satisfied by a direct delivery from the plant which has infinite supply. Note that this last option is equivalent to modeling a lost demand or using a source outside the system.

We analyze the system in two steps. First, we construct an aggregate model that enables us to calculate the fraction of demand that is satisfied by a direct delivery from the central warehouse and the fraction of demand that is satisfied by a direct delivery from the plant. These fractions are identical for all locals, even if they have different demand rates and stock levels.

Second, we construct a model for every local separately that enables us to estimate the fraction of demand satisfied by stock on hand and the fraction of demand satisfied by ELT. In this step we apply a technique introduced by Axsäter [2].

The literature on multi-echelon inventory systems for service parts covers over 25 years of research. The METRIC model of Sherbrooke [10] is widely considered as the first multi-echelon inventory model for service parts. The METRIC model is capable of determining the optimal stock levels that minimize the expected backorders at the locals subject to a budget constraint. Although METRIC is not an exact model, it gives good results and has been applied in practice by many companies. An improved version of METRIC called VARI-METRIC is presented by Graves [4]. Muckstadt and Thomas [7] extended the METRIC model with the option of direct deliveries from the central warehouse or the plant in case of a stockout situation at the local warehouse. They observe that in practice most multi-echelon systems are managed using adaptations of single location models. Their main goal is to show that such models can be dramatically inferior to models that take advantage of the system's structure. However, they do not investigate explicitly in their multi-echelon model the trade off between costs and service performance when using direct deliveries from a higher echelon. They also don't allow for lateral transhipments between the local warehouses. Moinzadeh and Schmidt [6] investigate the use of emergency replenishments for a single-echelon model with deterministic lead times. An emergency replenishment is issued when the stock level drops below a certain threshold value and the expected arrival time of the first pipeline order exceeds the emergency transhipment time. They compare their policy with a number of other policies. Aggarwal and Moinzadeh [1] conduct a similar policy evaluation study for a two-echelon system where the locals can issue an emergency replenishment when the number of outstanding orders drops below a certain threshold value. They do not take into account the pipeline orders and they do not allow for lateral transhipments between the bases. The central location is represented by a production plant that produces to order and therefore has no stock on hand. The production plant is modelled as an $M|G|1$ waiting queue where emergency orders have priority over normal replenishment orders.

Lee [5] presents a two-echelon model with one-for-one replenishment in which he allows for lateral transhipments between the local warehouses. The local warehouses are supplied from a central warehouse which in turn is supplied from the plant which is assumed to have infinite supply. The local warehouses are grouped into a number of pooling groups. Within each group

the warehouses are assumed to be identical. If demand cannot be satisfied from stock on hand, ELT's are used to fill the demand from another warehouse in the same pooling group that has stock on hand. If this is not possible, the demand is backordered. Lee derives approximate expressions for the fraction of demand satisfied from stock on hand, the fraction of demand that is satisfied by ELT, and the fraction of demand that is backordered. He compares his approximations with simulation results when different sourcing rules (which local warehouse in the group will source the ELT?) are used. The results show that the differences between sourcing rules are not significant and that the approximation is accurate for high values of the fill rate (> 0.70). In his paper Lee also presents an algorithm for determining optimal stocking levels such that costs (holding, backorder, and ELT) are minimized, subject to service level constraints.

Axsäter [2] analyzes the same system as Lee does. The local warehouses in each pooling group do not have to be identical. He uses a different modeling approach by concentrating on the demand processes at the local warehouses. When stock on hand is positive, the demand faced by the local warehouse equals the normal demand plus some ELT demand from other local warehouses in the same pooling group. When stock on hand is not positive, the only demand faced by the local warehouse is the backordered demand. Steady-state probabilities are derived by assuming exponentially distributed replenishment times. The analytical results are compared with simulation results. For the case with identical warehouses in each pooling group Axsäter compares his model with Lee's model and finds better results. Also for the case with nonidentical warehouses Axsäter's model gives satisfactory results.

Sherbrooke [11] presents a simulation study in which he investigates the added value of using lateral transhipments in a two-echelon depot-base system for repairable items. In contrast with Lee and Axsäter, Sherbrooke allows for delayed lateral transhipments. This means that if a base has zero inventory and receives a replenishment order from the depot, this unit may be laterally transhipped to another base with a backorder. Sherbrooke assumes that an ELT, normal or delayed, is only issued if it will arrive sooner than a pipeline unit. Upper and lower bounds for the expected system backorders are derived. Next regression analysis on the simulation data is used to derive approximate expressions for the expected system backorders. For depot-only-repairable items Sherbrooke shows that an average backorder reduction of 30–50% is possible (with a maximum of 72%) when using ELT's.

Pyke [9] presents a simulation study for a two-echelon system for repairable parts for electronic equipment on military aircraft. His main goal is to investigate the use of priority rules for the central repair shop in conjunction with priority rules for allocating repaired items to the bases. With regard to lateral transhipments he concludes that the improvement of the performance is marginal when decreasing the lateral transhipment times. The major gain is obtained in the limit, when the lateral transhipment times go to zero.

Dada [3] models a two-echelon system with priority shipments which is closely related to our model. Demand that can not be satisfied from stock on hand at a local warehouse is satisfied through emergency lateral transhipments or a direct delivery from the central warehouse. If this is not possible, Dada assumes that any item in transit from the central warehouse to the local warehouses can be used to satisfy this demand. Therefore Dada assumes that full information is available about items in transit and that these items can be redirected to any other destination after arrival at the original destination. Although current information systems make this sort of pipeline information more accessible, it is in most practical situations not feasible to apply this option. Think for example of boats, planned production situations or trucks of service providers. We therefore restrict ourselves to priority shipment options from physical stock locations. Dada applies a similar modeling approach in which he first constructs an aggregate model for the local warehouses and next presents a disaggregation scheme to find the performance of the individual locals. However, Dada assumes in his model that all local warehouses have identical lead times

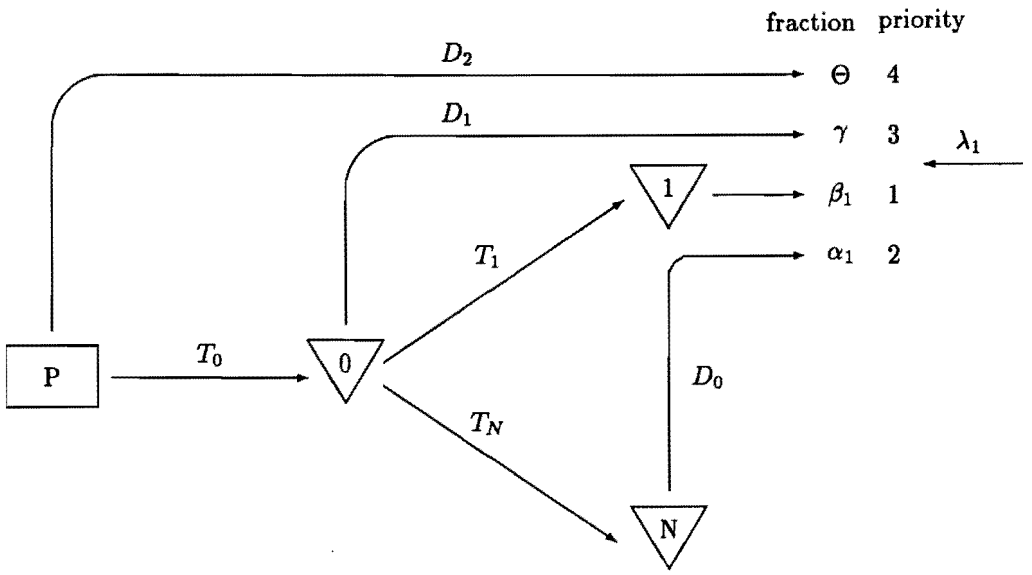

Figure 1: The inventory system

and stock level one. The analysis of his model is rather complex and the approximation scheme he presents for the case of non-identical locals does not always converge.

This paper is organised as follows. In section 2 we introduce our model and list the assumptions and notations. In section 3 we validate our model by means of simulation. Here we also present a sensitivity analysis for the lead time distribution. In section 4 we introduce a cost function that we use to find optimal stock levels. The use of emergency supply flexibility in a distribution network for service parts is associated with certain costs. We compare the cost results of our model with the cost results of the standard VARI-METRIC model which does not include any emergency supply flexibility. Finally, in section 5 we give some concluding remarks and topics for further research.

2 Model description

2.1 Assumptions and notation

As presented earlier, our inventory model consists of a number, N, of local warehouses and one central warehouse. We will use index i, $i = 1, \ldots, N$ to denote a specific local warehouse, and 0 for the central warehouse. Customers are assumed to arrive at local warehouse i according to a Poisson process with constant intensity λ_i. Furthermore, we let $\bar{\lambda}$ denote the total arrival intensity of customers, i.e., $\bar{\lambda} = \sum_{i=1}^{N} \lambda_i$. The structure of the inventory system and the policy used for filling demand is depicted in figure 1.

We let S_i denote the stock level at local warehouse i, and S_0 the stock level at the central warehouse. A customer arriving at local warehouse i will receive an item from stock on hand if stock is available. By β_i we denote the fraction of the demand λ_i that can be met directly from stock on hand. Since one-for-one replenishment is used, a customer served from stock on hand will trigger a replenishment order from the central warehouse to local warehouse i.

If a customer arrives at local warehouse i when this warehouse is out of stock, we serve the customer by issuing an ELT from a randomly chosen neighbor (local warehouse) with stock on hand. The fraction of the demand λ_i that is met by ELT is denoted by α_i. In principle, this could

be seen as a redirection of the customer to another local warehouse. In particular, the central warehouse will receive an order for replenishment from the local warehouse sourcing the ELT. We assume that the customer initiating the lateral transhipment will wait for this item, although the local warehouse could receive items through normal replenishment while the customer is waiting. Note that the local warehouses are assumed to form one pooling group, and that the average lateral transhipment time, D_0, is identical for all transhipments between local warehouses.

In the case when all local warehouses are out of stock, the central warehouse has stock on hand, and a customer arrives at any local warehouse, the customer is served through direct delivery from the central warehouse. By γ we denote the fraction of the total demand $\bar{\lambda}$ that is met in this way. It is clear from our assumptions that γ is also the fraction of customers arriving at local warehouse i that will be served through direct delivery from the central warehouse. As above, we assume that a customer initiating a direct delivery from the central warehouse will wait for this item to arrive, with an average direct delivery time of D_1.

If a customer arrives when there is no stock on hand, locally or at the central warehouse, the customer is served by direct delivery from the plant. The fraction of the total demand $\bar{\lambda}$ that will be met in this way is denoted by Θ. This is also the fraction of customers arriving at any local warehouse that will be satisfied through direct delivery from the plant. Again, we assume that a customer initiating a direct delivery from the plant will wait for this item to arrive, with an average direct delivery time of D_2.

If the central warehouse is out of stock when receiving a replenishment order, the demand is backlogged. Demand, including backlogged demand, at the central warehouse is satisfied on the basis of first come first served (FCFS). For local warehouse i, the time between placing a replenishment order and receiving the ordered item is called the local replenishment leadtime. It consists of the transportation time from the central warehouse to local warehouse i plus, in case of central stock-out, an additional waiting time for a spare to become available. Since our approach will be based on Markov analysis, we assume that the local replenishment leadtimes are independent exponentially distributed with mean L_i at local warehouse i. The average shipment time is denoted T_i and the expected delay denoted Δ, and thus $L_i = T_i + \Delta$.

N :	number of local warehouses
λ_i :	customer arrival rate at local warehouse i
$\bar{\lambda}$:	total customer arrival rate at all local warehouses
S_0 :	stock level at the central warehouse
S_i :	stock level at local warehouse i
T_0 :	shipment time from the plant to the central warehouse
T_i :	shipment time from the central warehouse to local warehouse i
Δ :	delay at the central warehouse
L_i :	local replenishment leadtime at local warehouse i
D_0 :	ELT time between local warehouses
D_1 :	direct delivery time from the central warehouse
D_2 :	direct delivery time from the plant
α_i :	fraction of demand at local warehouse i satisfied through ELT
β_i :	fraction of demand at local warehouse i satisfied from stock on hand
γ :	fraction of total demand satisfied through direct delivery from the central warehouse
Θ :	fraction of total demand satisfied through direct delivery from the plant

Table 1: List of notation

When a customer is served by stock on hand, an ELT or a direct delivery from the central warehouse, this will also result in demand at the central warehouse. Consequently, the central warehouse issues a replenishment order to the plant. The expected time the central warehouse has to wait before it receives the ordered item, we call the central replenishment leadtime. As mentioned earlier, we assume that the plant has infinite supply, why the central replenishment leadtime equals the shipment time from the plant to the central warehouse, T_0. In our model, we assume that the shipment times are independent and exponentially distributed with mean T_0. A summary of the notation used can be found in table 1.

2.2 Aggregate model for γ and Θ

As mentioned earlier, we will use a two-step procedure to find estimates of γ, Θ, and α_i and β_i for all locals i. As a first step we find estimates of γ, Θ, and Δ, the expected delay at the central warehouse which we will need in the second step as described in section 2.3. The approach, which follows closely the idea described in Dada [3], is based on the observation that from the central warehouse's point of view, the local warehouses will behave as one aggregate warehouse with stock level $\bar{S} = \sum_{i=1}^{N} S_i$. The idea is then to construct a finite two-state, (j,k), Markov model where

j = stock on hand at central warehouse, $-\bar{S} \leq j \leq S_0$
k = stock on hand at the aggregate local warehouse, $0 \leq k \leq \bar{S}$

In this case, there are three events leading to a change of state: A customer arrives at the aggregate local warehouse, a replenishment order arrives at the aggregate local warehouse, or a replenishment order arrives at the central warehouse. The rate at which customers arrive at the aggregate local warehouse is $\bar{\lambda}$. The other two rates are denoted

$\bar{\mu}$ = the rate at which a replenishment order arrives at the aggregate local warehouse,
μ_0 = the rate at which a replenishment order arrives at the central warehouse.

It is clear that $\mu_0 = 1/T_0$, but the rate $\bar{\mu}$ deserves some further attention. If the shipment times are all exponentially distributed with the same mean, i.e. $T_i = T_1 \ \forall i$, then $\bar{\mu} = 1/T_1$, and the model is exact. If this is not the case, we use the following approximation: $\bar{T} = \sum_{i=1}^{N} \lambda_i T_i / \bar{\lambda}$ and $\bar{\mu} = 1/\bar{T}$. That is, we assume that the rate at which items are being sent from the central warehouse to local i is equal to the demand rate λ_i. Due to lateral transhipments, this does not have to be the case. Moreover, we assume that the shipment time from central warehouse to the aggregate local is exponentially distributed with mean $\bar{T}$. Note that even if all individual shipment times are exponentially distributed but with different means, this will not be the case.

The state space and corresponding transition rates are depicted in figure 2 for the case when $S_0 = 2$ and $\bar{S} = 2$. From this picture it is clear how to form the state space for different values of S_0 and $\bar{S}$. By solving the corresponding linear equation system together with the normalizing constraint $\sum_{j=-\bar{S}}^{S_0} \sum_{k=0}^{\bar{S}} \pi_{jk} = 1$ we find the steady-state probabilities

$\pi_{jk} = \mathrm{P}[j$ items on hand at the central and k items on hand at the aggregate local warehouse$]$

A negative value of j corresponds to $-j$ backorders at the central warehouse. Having found the steady-state probabilities we can find γ, Θ and Δ as follows:

$$\gamma = \sum_{j=1}^{S_0} \pi_{j0}, \quad \Theta = \sum_{j=-\bar{S}}^{0} \pi_{j0}, \quad \Delta = \frac{1}{(1-\Theta)\bar{\lambda}} \sum_{j=-\bar{S}}^{-1} (-j) \sum_{k=0}^{\bar{S}+j} \pi_{jk}$$

The expression for Δ is found by applying the well-known result by Little which states that the average waiting-time is equal to the expected number of backorders divided by the arrival rate.

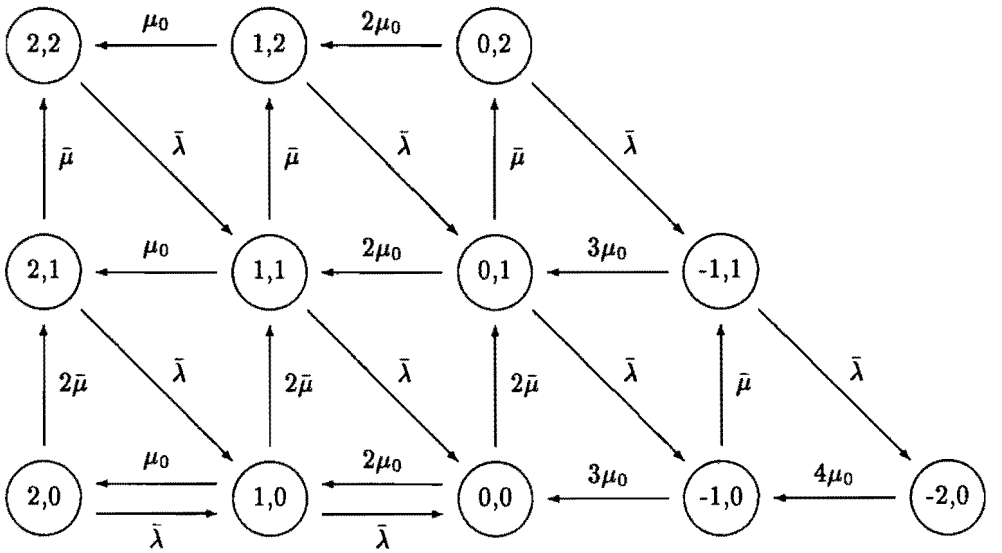

Figure 2: Aggregate state space model

2.3 Model for α_i and β_i

The objective of the second step of our heuristic is to find estimates of α_i and β_i for all locals i. Again we will use Markov analysis, following closely the approach described by Axsäter [2]. The main idea is to adjust the demand rate at a local warehouse by taking into consideration that the local warehouse will sometimes be used as a source for ELT's to other locals.

When local warehouse i has stock on hand, it will face the regular demand λ_i, and ELT's from other locals with an average rate of e_i. (The actual expression for e_i is presented later.) When the inventory position is zero the demand is zero, since customers arriving in this state are satisfied either by an ELT or a direct delivery. Let $g_i = \lambda_i + e_i$ denote the adjusted demand rate at local i. We make the approximation that the demand process at local i is Poisson with rate g_i, and that the demand processes at different locals are independent.

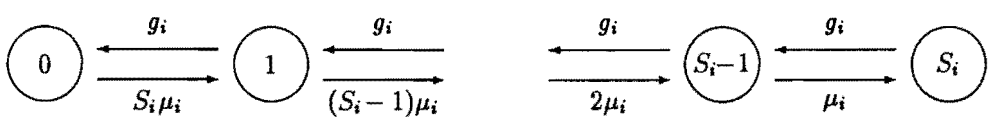

Figure 3: State space model for local i

As mentioned previously, we assume that the replenishment leadtimes for local warehouse i are independent and exponentially distributed with mean $L_i = T_i + \Delta$. The replenishment rate is then $\mu_i = 1/L_i$ and the corresponding state space is depicted in figure 3. The steady-state equations can be solved analytically. Let

$$p_j^i = \mathrm{P}[j \text{ items on hand at local warehouse } i], \quad 0 \le j \le S_i.$$

Then,

$$p_{S_i}^i = \left\{ \sum_{j=0}^{S_i} \frac{(g_i/\mu_i)^j}{j!} \right\}^{-1} \quad \text{and} \quad p_j^i = \frac{(g_i/\mu_i)^{S_i-j}}{(S_i-j)!} p_{S_i}^i$$

After computing the steady-state probabilities, we can readily find the estimates $\beta_i = 1 - p_0^i$ and $\alpha_i = 1 - \beta_i - \gamma - \theta$.

Still remaining is the expression for e_i. In the case when all local warehouses are identical, and a randomly chosen neighbor is used to source an ELT, $e_i = \alpha_i \lambda_i / \beta_i$. Note that e_i depends on α_i and β_i. Hence, we use an iterative procedure where we alternately update the values for β_i, α_i, and e_i. However, as Axsäter [2] also notes, convergence is obtained after a few iterations. In our case, we start by assuming $\beta_i = 1 - \gamma - \theta$ and $\alpha_i = 0$, which implies that e_i is initially zero for all i.

When the local warehouses are not identical, the expressions for e_i become more complicated, but the general idea remains the same, i.e, we iteratively solve for α_i, β_i, and e_i. If a randomly chosen neighbor is used to source an ELT, the general expression for e_i looks as follows:

$$e_i = \sum_{k \neq i} \frac{\alpha_k \lambda_k}{(1 - \prod_{j \neq k}(1 - \beta_j))} \sum_{\substack{v = (v_j) \\ j \neq i,k \\ v_j \in \{0,1\}}} \frac{\prod \beta_j^{v_j} (1 - \beta_j)^{1 - v_j}}{1 + \sum v_j}$$

The second sum is taken over all zero-one vectors of length $N - 2$, where the components are numbered $1, \ldots, j-1, j+1, \ldots, k-1, k+1, \ldots, N$. If $N = 2$, this second sum is defined to be equal to one, which gives

$$e_1 = \alpha_2 \lambda_2 / \beta_1 \quad \text{and} \quad e_2 = \alpha_1 \lambda_1 / \beta_2.$$

When $N = 3$,

$$e_1 = \alpha_2 \lambda_2 (1 - \beta_3/2)/(\beta_1 + \beta_3 - \beta_1 \beta_3) + \alpha_3 \lambda_3 (1 - \beta_2/2)/(\beta_1 + \beta_2 - \beta_1 \beta_2),$$

and analogously for $i = 2, 3$.

The formulas for e_i presented above were based on a random choice of source for an ELT. However, we would like to point out that the modeling approach could be used also in the case when each local uses a priority list for determining the source of ELTs. If the local warehouse looks to all other locals before using a direct delivery from the central warehouse or plant, the aggregate model we presented in subsection 2.2 would still apply. In particular, the values for γ and Θ would again be exact if the shipment times are exponentially distributed with the same mean. The difference is that the formulas for e_i change, and thus the values for α_i and β_i.

3 Model validation

In this section, we present simulation results in order to show the accuracy of our model. We consider a situation with three local warehouses, i.e., $N = 3$ throughout this section. For each simulation run we simulated a minimum of 500,000 thousand customer arrivals at each local warehouse. We also test the sensitivity of our model with regard to the leadtime distribution.

3.1 Simulation results for α_i, β_i, γ and θ

Since our modeling technique assumes exponential leadtimes, we first consider a benevolent situation when the central leadtime and the shipment times are drawn from exponential distributions with mean T_0 and T_i respectively.

First, we consider a set of 14 problems for which the three local warehouses are identical. Due to direct deliveries, the results from our model can not be compared directly to, e.g., the models of Lee [5] and Axsäter [2]. However, the parameter values used in the problems are chosen similar

case	λ_i	S_i	S_0	γ			Θ			β_i		
				ESM	exp	det	ESM	exp	det	ESM	exp	det
1	0.02	1	1	0.00	0.00	0.00	0.02	0.02	0.02	0.84	0.84	0.84
2			2	0.00	0.00	0.00	0.01	0.01	0.00	0.92	0.92	0.92
3	0.06	1	1	0.00	0.00	0.00	0.23	0.23	0.23	0.48	0.48	0.48
4			2	0.00	0.00	0.00	0.13	0.13	0.13	0.61	0.61	0.61
5		2	1	0.00	0.00	0.00	0.03	0.03	0.03	0.83	0.81	0.81
6			2	0.00	0.00	0.00	0.01	0.01	0.01	0.91	0.89	0.89
7	0.10	1	2	0.00	0.00	0.00	0.32	0.32	0.32	0.40	0.40	0.40
8			3	0.00	0.00	0.00	0.23	0.23	0.22	0.49	0.49	0.49
9			6	0.02	0.02	0.02	0.06	0.06	0.06	0.67	0.67	0.67
10			10	0.05	0.05	0.05	0.00	0.00	0.00	0.71	0.71	0.71
11		2	2	0.00	0.00	0.00	0.09	0.09	0.09	0.71	0.69	0.69
12			3	0.00	0.00	0.00	0.05	0.05	0.05	0.80	0.78	0.78
13		3	2	0.00	0.00	0.00	0.01	0.01	0.01	0.90	0.88	0.88
14			3	0.00	0.00	0.00	0.01	0.01	0.01	0.95	0.93	0.93

Table 2: Results for identical local warehouses, $N = 3$, $T_0 = 15$ and $T_i = 3$

to theirs. The central shipment time, T_0, equals 15, and the local shipment times, T_i, are all equal to 3. The results from the simulation with exponential distributions (exp) and from our emergency supply model (ESM) are shown in table 2 for γ, Θ and β_i. The value for α_i can readily be obtained from the relationship $\alpha_i + \beta_i + \gamma + \Theta = 1$. Note that when the shipment times are the same for all locals and independent exponentially distributed, the values for γ and Θ should be exact, which is indeed the case.

Second, we also consider a set of 18 problems for which the local warehouses are not identical, but have different demand rates λ_i, or different shipment times T_i. The results are shown in table 3 on page 10 and 11. Again, the γ and Θ values should be exact for the problems marked 'a', which they also are.

For the case with identical locals (table 2) our model gives excellent results. When the local stock levels are equal to one our model even yields exact results. For the case with non-identical locals (table 3) the model still performs very good.

3.2 Sensitivity to leadtime distribution

Our modeling assumptions include exponentially distributed shipment times for normal replenishments. However, in practice these shipment times are close to deterministic since they consist of transportation times between inventory locations. Therefore, we also simulated the system with deterministic leadtimes. The results are also presented in tables 2 and 3 (det).

An analysis of these results shows that the service performance of the system is almost identical for exponential and deterministic shipment times. Apparently the leadtime distribution does not affect the service performance. In fact, the key METRIC assumption is that Palm's theorem for infinite server queues applies to the replenishment process of the local warehouses as well. Palm's theorem [8] states that the distribution of parts in resupply is only dependent on the replenishment time distribution through its mean. The results indicate that our model is to a large extent insensitive to the choice of the leadtime distribution.

					γ			Θ			β_i		
case	λ_i	L_i	S_i	S_0	ESM	exp	det	ESM	exp	det	ESM	exp	det
	0.01	3	1	1	0.00	0.00	0.00	0.02	0.02	0.02	0.89	0.88	0.88
1a	0.02	3									0.84	0.84	0.84
	0.03	3									0.80	0.81	0.81
	0.01	3	1	2	0.00	0.00	0.00	0.01	0.01	0.00	0.95	0.94	0.94
2a	0.02	3									0.91	0.91	0.91
	0.03	3									0.88	0.89	0.89
	0.02	1	1	1	0.00	0.00	0.00	0.02	0.02	0.02	0.88	0.88	0.87
1b	0.02	3									0.84	0.84	0.84
	0.02	5									0.82	0.82	0.82
	0.02	1	1	2	0.00	0.00	0.00	0.01	0.01	0.00	0.95	0.95	0.95
2b	0.02	3									0.92	0.92	0.92
	0.02	5									0.88	0.88	0.88
	0.02	3	1	2	0.00	0.00	0.00	0.13	0.13	0.13	0.69	0.67	0.66
4a	0.06	3									0.61	0.61	0.61
	0.10	3									0.54	0.56	0.56
	0.02	3	2	1	0.00	0.00	0.00	0.03	0.03	0.03	0.92	0.88	0.88
5a	0.06	3									0.81	0.80	0.80
	0.10	3									0.72	0.73	0.73
	0.02	3	2	2	0.00	0.00	0.00	0.01	0.01	0.01	0.97	0.94	0.94
6a	0.06	3									0.90	0.89	0.89
	0.10	3									0.83	0.83	0.83
	0.06	1	1	2	0.00	0.00	0.00	0.13	0.13	0.13	0.68	0.68	0.67
4b	0.06	3									0.62	0.61	0.61
	0.06	5									0.57	0.57	0.57
	0.06	1	2	1	0.00	0.00	0.00	0.03	0.03	0.03	0.86	0.84	0.84
5b	0.06	3									0.83	0.81	0.81
	0.06	5									0.80	0.78	0.78
	0.06	1	2	2	0.00	0.00	0.00	0.01	0.01	0.01	0.94	0.92	0.92
6b	0.06	3									0.91	0.89	0.89
	0.06	5									0.88	0.87	0.87
	0.05	3	1	6	0.02	0.02	0.02	0.06	0.06	0.06	0.73	0.73	0.73
9a	0.10	3									0.67	0.67	0.67
	0.15	3									0.61	0.62	0.62
	0.05	3	1	10	0.05	0.05	0.05	0.00	0.00	0.00	0.77	0.77	0.77
10a	0.10	3									0.71	0.71	0.71
	0.15	3									0.66	0.66	0.66

Table 3: Results for nonidentical local warehouses, $N = 3$ and $T_0 = 15$

					γ			Θ			β_i		
case	λ_i	L_i	S_i	S_0	ESM	exp	det	ESM	exp	det	ESM	exp	det
	0.05	3	2	2	0.00	0.00	0.00	0.09	0.09	0.09	0.79	0.74	0.74
11a	0.10	3									0.70	0.68	0.68
	0.15	3									0.63	0.63	0.63
	0.05	3	2	3	0.00	0.00	0.00	0.05	0.05	0.05	0.88	0.83	0.83
12a	0.10	3									0.80	0.77	0.77
	0.15	3									0.72	0.72	0.72
	0.10	1	1	6	0.02	0.01	0.01	0.06	0.06	0.05	0.82	0.81	0.81
9b	0.10	3									0.69	0.68	0.68
	0.10	5									0.59	0.59	0.59
	0.10	1	1	10	0.05	0.03	0.03	0.00	0.00	0.00	0.88	0.87	0.87
10b	0.10	3									0.73	0.72	0.72
	0.10	5									0.63	0.62	0.62
	0.10	1	2	2	0.00	0.00	0.00	0.09	0.09	0.09	0.76	0.73	0.73
11b	0.10	3									0.71	0.69	0.69
	0.10	5									0.67	0.65	0.65
	0.10	1	2	3	0.00	0.00	0.00	0.05	0.05	0.05	0.86	0.82	0.82
12b	0.10	3									0.80	0.78	0.78
	0.10	5									0.76	0.74	0.74

Table 3: *cont'd*

4 Economic evaluation

4.1 Cost structure

In order to evaluate the performance of the model we need a cost structure that takes into account the different operational cost factors. These include: inventory holding costs, normal replenishment costs (to the central warehouse and to the local warehouses), emergency shipment costs (ELT, direct delivery from the central warehouse and direct delivery from the plant) and penalty costs for customers who have to wait. We need the following cost parameters to make an economic evaluation:

c : unit price of the item

h_0 : inventory holding cost at central warehouse per item per time unit expressed as a fraction of the unit price

h_i : inventory holding cost at local warehouse i per item per time unit expressed as a fraction of the unit price

r_0 : normal replenishment cost per item for the central warehouse

r_i : normal replenishment cost per item for local warehouse i

e_0 : emergency replenishment cost for using an ELT per item

e_1 : emergency replenishment cost for using a direct delivery from the central warehouse per item

e_2 : emergency replenishment cost for using a direct delivery from the plant per item

z_i : penalty cost per time unit for a waiting customer at local warehouse i

We assume that the cost for using a particular kind of emergency shipment is equal for all locals. However, it is possible to differentiate these costs for the different locals. The expected waiting time for an arbitrary customer at local i, w_i, can be expressed as follows:

$$w_i = \alpha_i D_0 + \gamma D_1 + \theta D_2$$

Given the steady-state probabilities π_{jk} and p_j^i we calculated in section 2, we can now formulate the total cost TC per time unit as follows:

$$\begin{aligned} \text{TC} &= c \cdot h_0 \sum_{j=1}^{S_0} \sum_{k=0}^{\bar{S}} j\pi_{jk} + c \sum_{i=1}^{N} h_i \sum_{j=1}^{S_i} j p_j^i + e_0 \sum_{i=1}^{N} \alpha_i \lambda_i + e_1 \gamma \bar{\lambda} + e_2 \theta \bar{\lambda} \\ &\quad + r_0(1-\theta)\bar{\lambda} + \sum_{i=1}^{N} r_i(1-\theta-\gamma)\bar{\lambda} + \sum_{i=1}^{N} \lambda_i w_i z_i \end{aligned}$$

We use this cost function to find the optimal stock levels that minimize the total costs.

4.2 Numerical evaluation

In this section we present, for different parameter values, the minimum total cost, TC*, and associated optimal stock levels, S_0^* and S_i^*, resulting from our model and cost structure. We assume that we have three identical local warehouses. Our main interest is to investigate the influence of certain parameters on the minimum total cost and optimal stock levels, namely: the demand rate (λ_i), the inventory holding cost fractions (h_0, h_i), the emergency replenishment costs (e_0, e_1, e_2) and the penalty cost (z_i). The remaining system parameters are fixed as follows (time units in days):

$$c = 10000, \quad r_0 = r_i = 10, \quad T_0 = 15, \quad T_i = 3, \quad D_0 = 0.5, \quad D_1 = 1, \quad D_2 = 2.$$

The parameters we vary in the experiment are set as follows:

$$\lambda_i \in \{0.02, 0.06, 0.10\}, h_i \in \{0.10/365, 0.30/365\},$$

$$e_{0,1,2} \in \{(10, 30, 100), (30, 90, 300)\}, z_i \in \{100, 1000\}.$$

The inventory holding costs are identical for all stocking locations and are 10% and 30% respectively of the unit price c per year. The results are shown in table 4.

4.3 Comparison with VARI-METRIC

Table 4 also shows the minimum total cost (TC^v) and the optimal stock levels (S_0^v, S_i^v) if no emergency supply flexibility exists in the system. The cost structure is the same as before except for emergency transportation costs that will not occur in this case. We calculated the relevant costs using the VARI-METRIC technique as described by Graves [4]. We see that in all the 24 cases we analyzed, the policy of using emergency supply flexibility as described in this paper results in a lower total cost. The right-most column of table 4 shows the relative decrease in costs when using emergency supply flexibility. A maximum cost reduction of 43.9% and a minimum of 13.2% is obtained. The results also show that in many cases the stock levels are lower when using emergency supply flexibility. Especially the central stock level shows a significant decrease.

λ_i	h_i	$e_{0,1,2}$	z_i	TC^*	S_0^*	S_i^*	TC^v	S_0^v	S_i^v	%
0.02	0.10/365	(10,30,100)	100	8.84	0	1	13.55	2	1	34.8
			1000	15.04	2	1	21.01	2	2	28.4
		(30,90,300)	100	10.03	0	1	13.55	2	1	26.0
			1000	15.19	2	1	21.01	2	2	27.7
	0.30/365	(10,30,100)	100	17.27	1	0	30.76	1	1	43.9
			1000	32.03	1	1	48.92	3	1	34.5
		(30,90,300)	100	20.99	0	1	30.76	1	1	31.8
			1000	32.43	1	1	48.92	3	1	33.7
0.06	0.10/365	(10,30,100)	100	17.22	1	2	23.28	3	2	26.0
			1000	24.86	3	2	31.34	5	2	20.7
		(30,90,300)	100	18.81	1	2	23.28	3	2	19.2
			1000	25.08	4	2	31.34	5	2	20.0
	0.30/365	(10,30,100)	100	30.97	1	1	46.94	4	1	34.0
			1000	54.72	2	2	73.90	5	2	26.0
		(30,90,300)	100	36.89	2	1	46.94	4	1	21.5
			1000	55.44	2	2	73.90	5	2	25.0
0.10	0.10/365	(10,30,100)	100	23.11	4	2	29.01	6	2	20.3
			1000	32.03	4	3	38.34	7	3	16.5
		(30,90,300)	100	25.18	5	2	29.01	6	2	13.2
			1000	32.21	5	3	38.34	7	3	16.0
	0.30/365	(10,30,100)	100	42.54	2	2	62.01	5	2	31.4
			1000	71.28	3	3	92.86	6	3	23.2
		(30,90,300)	100	48.76	3	2	62.01	5	2	21.4
			1000	71.91	3	3	92.86	6	3	22.6

Table 4: Cost evaluation

5 Concluding remarks

In this paper we presented a two-echelon inventory system with one-for-one replenishment in which we modelled a number of supply alternatives. Customers that arrive at the local warehouses in a stockout situation are not backordered but satisfied through an emergency lateral transhipment, a direct delivery from the central warehouse, or a direct delivery from the plant. We presented an approximate model that is solved in two steps. In the first step an exact aggregate model is developed that combines all local warehouses into one warehouse. In the second step we use an approximate model to calculate the service performance at the various local warehouses. The numerical results indicate that the performance of our model is very close to the simulation results. Another important observation is that the distribution of the shipment times has a negligible impact on the service performance. This is important since we have to assume exponentially distributed shipment times in our model analysis. In an economical analysis we compared the optimal cost results of our model with the VARI-METRIC cost results. In all cases we found major cost reductions, which indicate that using emergency supply flexibility in a distribution network for service parts can be very beneficial.

Some topics for further research could be the modeling of more than one pooling group and extending the model to more than two echelons. Another important research question is the comparison of different policies. We compared our model with the VARI-METRIC model in

which all excess demand is backordered. Comparisons with other policies in which some of the above emergency supply alternatives are used (or other alternatives?) can answer the question: When to use what kind of emergency supply flexibility in your distribution network?

References

[1] Aggarwal, P.K. and Moinzadeh K., "Order expedition in multi-echelon production/distribution systems", *IIE Transactions* 26 (1994) 86–96.

[2] Axsäter, S., "Modelling emergency lateral transhipments in inventory systems", *Management Science* 36 (1990) 1329–1338.

[3] Dada, M., "A two-echelon inventory system with priority shipments" *Management Science* 38 (1992) 1140–1153.

[4] Graves, S.C., "A multi-echelon inventory model for a repairable item with one-for-one replenishment", *Management Science* 31 (1985) 1247–1256.

[5] Lee, H.L., "A multi-echelon inventory model for repairable items with emergency lateral transhipments", *Management Science* 33 (1987) 1302–1316.

[6] Moinzadeh, K. and C.P. Schmidt, "An (S-1,S) inventory system with emergency orders", *Operations Research* 39 (1991) 308–321.

[7] Muckstadt, J.A. and L.J. Thomas, "Are Multi-Echelon Inventory Methods Worth Implementing in Systems with Low- Demand Rates?", *Management Science* 26 (1980) 483–494

[8] Palm, C., "Analysis of the Erlang Traffic Formula for Busy-Signal Arrangements", *Ericsson Technics* 5 (1938) 39–58.

[9] Pyke, D.F., "Priority repair and dispatch policies for repairable-item logistics systems", *Naval Research Logistics* 37 (1990) 1–30.

[10] Sherbrooke, C.C., "METRIC: A Multi-Echelon Technique for Recoverable Item Control", *Operations Research* 16 (1968) 122–141.

[11] Sherbrooke, C.C., "Multi-echelon inventory systems with lateral supply", *Naval Research Logistics* 39 (1992) 29–40.