TI 2010-025/4
Tinbergen Institute Discussion Paper

Service Parts Inventory Control with Lateral Transshipment that takes Time

Guangyuan Yang
Rommert Dekker

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at
http://www.tinbergen.nl.

Service Parts Inventory Control with Lateral Transshipments and Pipeline Stock Flexibility

Guangyuan Yang
Econometric and Tinbergen Institute, Erasmus School of Economics, Erasmus University Rotterdam, 3062 PA, Rotterdam. gyang@ese.eur.nl

Rommert Dekker
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, 3062 PA, Rotterdam. rdekker@ese.eur.nl

Adriana F. Gabor
Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, 3062 PA, Rotterdam. gabor@ese.eur.nl

Sven Axsäter
Department of Industrial Management and Logistics, Lund University, S-221 00, Lund. sven.axsater@iml.lth.se

In equipment-intensive industries such as truck, electronics, aircraft and dredging vessel manufacturing, service parts are often slow moving items for which the transshipment time is not negligible. However, this aspect is hardly considered in the existing service logistics literature. In this paper, we consider this aspect and propose a customer-oriented service measure which takes into account pipeline stock and lateral transshipment flexibility. We provide an approximation method for optimizing the stock allocation subject to this service measure. Via extensive numerical experiments, we show that our approximation performs very well with respect to both system performance and costs. Moreover, our numerical experiments indicate that including lateral transshipments and pipeline stock flexibility in inventory decisions is more beneficial than lateral transshipments alone. This effect is larger for high demand rates and high lateral transshipment costs. Results from a case study in the dredging industry confirm our findings. We therefore recommend introduction of pipeline stock information such as the track and trace information from freight carriers in existing ERP systems.

Key words: Customer-oriented Service Measure, Response Time, Lateral Transshipment Flexibility, Pipeline Stock Flexibility, Maritime Applications.

History: This paper was last revised on November 15, 2012.

1. Introduction

Research on service parts inventory control with lateral transshipments has been motivated by needs from various industries, including equipment-intensive industries such as truck, electronics, aircraft and dredging vessel manufacturing. Facing stochastic demand of critical service parts, a multi-location inventory control system often allows movement of stock between locations at the same echelon level or even across different levels in order to fulfil customers' demand in time. Many of these critical service parts are slow moving and heavy items for which, in some cases, air transport is impossible or prohibitively expensive. For example, in dredging industry, critical service parts usually weigh more than 1700 kg and therefore are way too costly to transport by air. The lateral transshipment time for these items can be longer than 3 weeks, and is not negligible compared to lead times (around 7 weeks).

Moreover, if transportation is slow, there may be considerable amounts of pipeline stock being transported between a production plant and local bases. In some cases, the average pipeline stock can be up to half of the average stock on hand. In these cases, it may be more profitable to wait till the pipeline stock arrives than to order via lateral transshipments. As a result, the timing of transshipments and normal replenishments becomes an important factor in decision making. To the best of our knowledge, this aspect is not much considered in the existing service parts literature.

Good customer-oriented performance measures are also lacking in the literature. The standard service levels, such as fill rates, are supplier-oriented; whereas customers only observe deliveries with no delays and deliveries within a certain response time in case of delays. Some studies (Kutanoglu and Mahajan, 2009) introduce more customer-oriented service levels by distinguishing the availability of items from different sources with different response times. However, since these studies ignore the fact that the pipeline stocks may arrive and be delivered to customers sooner than other emergency shipments, they still emphasize the operational processes of service suppliers.

Inventory sharing by lateral transshipments between local bases makes stock more valuable as the stock may be used by different bases. The benefits are clear when there is an agreement with a customer on the response time and lateral transshipment times are negligible. On the other hand, if transshipment times are not negligible, lateral transshipments could be detrimental for service levels, as products spend more time in transportation before reaching customers. Hence, the total system cost may be higher in this case due to higher requirements of base stocks.

Statement of contribution. The contribution of this paper is as follows: First, we propose a customer-oriented performance measure where both pipeline stock and lateral transshipment flexibility contribute to service performance. Second, we provide an approximation method for optimizing stock levels subject to this measure. The quality of this approximation and the benefits of lateral transshipment and pipeline stock flexibility are assessed via extensive numerical experiments. Based on these experixments, we find that the use of pipeline stock information improves the performance and costs of systems with lateral transshipments. We conclude that including lateral transshipment and pipeline stock flexibility in inventory decisions is more beneficial than lateral transshipments alone. Subsequently, we apply our method in a case study from dredging industry and confirm our findings in practice. Finally, we offer managerial insights on the lateral transshipment decisions when the transshipment time is non-negligible.

The paper is organized as the following. In Section 2, we review the inventory control literature on lateral transshipments. In Section 3 we present the inventory model. In Section 4 and 5, we give an approximation for the customer-oriented service measure in the context of a single-echelon inventory system with and without lateral transshipments between the local bases. In Section 6 we minimize the average inventory cost subject to service level constraints. We validate our methods and we assess the benefits of lateral transshipment and pipeline stock flexibility via extensive numerical experiments in Section 7. In Section 8, we apply our method in a case study for a global market leader in the dredging industry. In the last part, we draw our conclusion and offer managerial insights.

2. Literature review

In the past decades, a considerable amount of research has been dedicated to service parts inventory control with lateral transshipments. Paterson et al. (2011) provide an extensive literature review on lateral transshipments. Based on the timing of transshipments, they categorize the research into proactive and reactive transshipments. The research on reactive transshipments is divided into two categories: one with centralized systems, the other with decentralized systems. Most models with centralized systems (Lee, 1987; Axsäter, 1990; Alfredsson and Verrijdt, 1999; Diks and De Kok, 1996; Banerjee et al., 2003; Burton and Banerjee, 2005) assume that transshipment times are negligible, and find that lateral transshipments improve the system performance.

In Lee (1987), a two echelon inventory system with continuous review base stock policy, identical bases and negligible transshipment times is analysed. Demand occurs when there is a failure of a critical part, and is assumed to follow a Poisson process. Failed parts are replaced by stock on hand or lateral transshipments in case of a stock-out. The portion of demand met by stock on hand and the portion of demand met by lateral transshipments are evaluated based on three selection rules for the source base: random selection, maximum stock on hand, and smallest number of outstanding orders. No significant difference in the performance of the three rules is found. The paper concludes that lateral transshipments lead to substantial cost savings because less base stocks are needed at the bases.

Axsäter (1990) relaxes the restrictive assumption of identical bases and presents improved methods for approximating service levels. Alfredsson and Verrijdt (1999) extend Axsäter's model by allowing emergency shipments from a central warehouse and emergency shipments from a manufacturing facility such that no demand is back-ordered. They find that using both lateral transshipment and direct shipment flexibility results in significant cost reductions compared to using no supply flexibility at all. Simulation studies with negligible transshipment time, conducted by Banerjee et al. (2003) and Burton & Banerjee (2005), show that a policy with lateral transshipments is superior to one without lateral transshipments if the benefits of avoiding retail level shortages outweigh the additional delivery costs resulting from transshipments.

On the other hand, some of the relatively few recent studies (Grahovac and Chakravarty, 2001; Tagaras and Vlachos, 2002; Wong et al., 2005; Kutanoglu and Mahajan, 2009) consider non-negligible lateral transshipment times in their models.

Grahovac and Chakravarty (2001) study the benefits of lateral transshipments by comparing overall transportation, inventory holding, and customer waiting time costs in cases with transshipments and without lateral transshipments. They extend the model in Axsäter (1990) by assuming non-negligible transportation times, which are identical for emergency lateral transshipments between retailers and direct emergency orders from the distribution center. They find that, in a centralized supply chain, lateral transshipments often reduce the overall costs. They explain that lateral transshipments make "front-line" inventories more valuable, leading to stock levels larger or equal to the levels without lateral transshipments at retailers. On the other hand, lateral transshipments make inventory at the distribution center less valuable, leading to a stock level smaller or equal to the level without lateral transshipments. These two opposite trends may lead to higher stock levels in the system with lateral transshipments than without lateral transshipments.

To investigate the operational characteristics of lateral transshipments, Tagaras and Vlachos (2002) conduct extensive experiments with 5 demand distributions and non-negligible lateral transshipment times, and conclude that the benefits of risk pooling are substantial only when demand is highly variable. Moreover, they find that the effectiveness of lateral transshipments is superior for identical bases in a pooling group. This is even more pronounced when the lead times are relatively long and the demand is more variable.

Wong et al. (2005) study repairable service parts pooling in a multi-hub system for the airline industry. They include delayed lateral transshipments in their system performance approximation and optimal service parts stocking level determination. Regarding the choice of the source for lateral transshipments, they use the closest neighbour rule as it is more acceptable in practice than the random selection rule used by Axsäter (1990) and Alfredsson and Verrijdt (1999). They find that significant cost savings can be achieved by pooling the service parts inventories via lateral transshipments.

Kutanoglu and Mahajan (2009) point out that the most commonly used service measure, the fill rate, does not capture the time necessary to satisfy the demand. They introduce a time-based service level, i.e., the proportion of total demand satisfied within a specified time window. However, these service levels ignore the pipeline stocks that may arrive and be delivered to customers sooner than lateral transshipment from other local warehouses. The authors find the optimal stock levels subject to the time based service level constraint by enumerating over all possible stock profiles (stock levels across all local warehouses). They also conclude that lateral transshipments improve the inventory system performance.

The use of pipeline stock information has not been studied much in the literature. Axsäter (2003) designed a heuristic decision rule for lateral transshipments that takes into account the remaining delivery time of outstanding orders. He assumes that each base follows an (R, Q) inventory policy and lateral transshipment times are negligible. In this paper, bases follow a base stock policy. We use the pipeline information not only in deciding the lateral transshipment rule, but also when calculating the customer oriented service level, i.e., the proportion of customers that are served within time T by stock on hand, pipeline stock or lateral transshipments.

Howard et al. (2010) studied the optimal time a customer should wait for pipeline stock at a base, when all bases follow a base stock policy. If no order arrives within this time, an emergency shipment from the warehouse takes place. Our paper differs from Howard et al. (2010) in several ways. In our case, a demand can also be satisfied by lateral transshipments, as long as this can be done within the prespecified time limit T. We do not optimize the value of T, and assume that all bases have the same value of T. Demand that cannot be satisfied within time T is in our case back ordered. Moreover, in our analysis, we use a different queuing model.

3. Model description

We consider a single item inventory model with one or two echelons. The single-echelon network consists of a production plant and a number of local bases, which operate as follows.

- The production plant replenishes the local bases within a base specific replenishment lead time. Furthermore, the plant has ample capacity, i.e. the replenishment lead time is not affected by the number of outstanding replenishment orders at the plant.

- Each customer is assigned to the closest local base. All demands at local bases and replenishment orders at the plant are fulfilled according to a first-come, first-served policy.

- The service parts have no lost sales. If there is stock available at the local base, i.e. the inventory level at the local base is positive, the demand is fulfilled instantaneously. If there are no stocks available at the local base, i.e., zero or negative inventory level at the local base, the customer will wait for the part and the request will be satisfied later either from the pipeline stock or by lateral transshipments from prioritized neighbouring bases which have stock on hand.

- As key performance indicator, we use a customer-oriented service measure defined as the proportion of demand satisfied within a certain response time, either by stock on hand, by pipeline stocks or via lateral transshipments. In our model, two response times are considered, i.e., 0 and a pre-set time T. If demand is satisfied within 0 response time, it is fulfilled by stock on hand instantaneously. If demand is satisfied within response time T, it can be fulfilled by all the three sources. Hence, the second service level is always higher than the first one. There is a target service level for each of the response times, according to service agreements with the customer. Note that both service levels are important. The instantaneous service level prevents the service provider from postponing service in order to reduce costs and assures that the customer is satisfied most of the times. The service level within T on the other hand, models the flexibility in service the customer is willing to accept.

This model can be easily adapted to other inventory systems, as will be discussed in Section 8.

3.1. Assumptions and notations

Inventory control network and policy. The network has a set B of N local bases. The local bases follow a base stock $(S-1, S)$ policy under continuous review. Each local base i has base stock S_i, $i = 1, \cdots, N$, which is replenished on a one-for-one, first-come-first served basis by the plant. This base stock inventory control system is very common for service parts, due to the high price and low demand characteristics of many of these items (Sherbrooke, 1968; Alfredsson and Verrijdt, 1999).

Customer demand. We assume that demand processes at the local bases are independent Poisson processes. The demand process at local base i will be denoted by $D_i(t)$ and has constant rate λ_i.

Lead times. The local replenishment lead time for base i is L_i. The lead times are assumed to be constant.

Operational rules. We specify the operational rules based on our observations from maritime industries, where lateral transsshipment times are usually much longer than in other industries. The inventory system is owned by a single principal, hence, inventory decisions regarding the allocation of stocks and the source to fulfill demands are taken centrally.

We assume that service at each base is offered based on a first-come first-served rule. Each base i has an associated ordered list B_i of other bases from which lateral transsshipment takes less than T time units. In practice, the list B_i is often ordered in increasing order of the distance to i.

Suppose a demand request arrives at base i. If this demand can be satisfied by stock on hand the demand is fullfilled and a replenishment order is issued. If there is no stock on hand at i, one first checks whether the demand can be fulfilled by replenishment orders (pipeline stock) that arrive within time T at base i. If yes, the demand will be fulfilled by base i and a replenishment order is issued. If there is no stock available within time T at base i, one checks the bases in B_i in the prescribed order. If there is a base in B_i with stock on hand, the demand will be satisfied by the first base in the list with stock. If no base in B_i has stock on hand, the demand is backordered and will be satisfied by base i, in a time longer than T.

Note that a base $i \in B$ faces two types of demand: *direct demand*, i.e., the requests that are originally issued for base i, and *lateral transsshipment requests*, i.e., requests that originally arrived at another base which faces stock out. The major difference between direct demands and lateral transsshipment requests is that for the former both the stock on hand and the pipeline stock are checked, whereas for the latter only the stock on hand is considered. Moreover, note that the demand requests arriving at i can be satisfied by stock on hand at i, or pipeline stock that arrives at i within T, by lateral transsshipment from bases in B_i or by pipeline stock that arrives at i after time T.

Service performance measures. The service performance of the network is measured by the customer-oriented service level with two key performance indicators: instantaneous service level (SL^0) and Service Level within response time T (SL^T). SL^0 is the proportion of demand satisfied within response time 0, i.e., instantaneously. SL^T is the proportion of demand satisfied within response time T, including the demands fulfilled instantaneously and the demands fulfilled by lateral transshipments from other bases. There is a target for the service performance according to service level agreements with customers. The service levels are required to be above given targets, i.e., $SL^0 \geq \phi$ and $SL^T \geq \tau$, where $0 < \phi < \tau < 1$.

We conclude this section by the list of parameters that will be used throughout the paper.

List of parameters

λ_i $=$ arrival rate at base i

IL_i $=$ inventory level at base i

SL_i^0 $=$ proportion of direct demand at base i satisfied by stock on hand at base i

SL^0 $=$ proportion of total demand satisfied by stock on hand

SL_i^T $=$ proportion of direct demand of base i satisfied within time T by base i or other bases in B_i

SL^T $=$ proportion of total demand satisfied from stock on hand or by lateral transshipments

ω_i $=$ proportion of demand at base i satisfied from pipeline within time T

α_{ij} $=$ proportion of demand at base i satisfied by stock on hand at base j

θ_i $=$ proportion of demand at base i satisfied at base i after time T

Note that for a base i, $SL_i^0 + \omega_i + \sum_{j \in B_i} \alpha_{ij} + \theta_i = 1$ and that $SL_i^T = 1 - \theta_i$.

4. Evaluation of performance without lateral transshipments

In this section we analyse the model where all the demand at a base is either satisfied by stock on hand or via pipeline stock; no lateral transshipments are possible. For this model, we calculate the instantaneous service level and the service level within time T.

Instantaneous service level. For an $(S_i - 1, S_i)$ inventory system in equilibrium, the inventory level at base i satisfies $IL_i = S_i - N_i$, where N_i is the number of replenishment orders in the pipeline. Note that, since a replenishment order is placed at the arrival of each customer, N_i can be viewed as the number of busy servers in an $M/D/\infty$ system, which by Palm's theorem (Palm, 1938), is distributed in steady state as a Poisson variable with parameter equal to the product of the arrival rate and mean service time. Consequently, for each $j \leq S_i$,

$$P(IL_i = j) = po(S_i - j; \lambda_i L_i).$$

where $po(k;\beta) = e^{-\beta}\frac{\beta^k}{k!}$, $k \in \mathbf{N}$ denotes the probability mass function of a Poisson variable with rate β. Thus, by Pasta, the fraction of demand that can be satisfied immediately from stock on hand SL_i^0 is equivalent to the fraction of time with positive stock on hand (see Axsäter, 2006). Hence,

$$SL_i^0 = P(IL_i > 0) = Po(S_i - 1; \lambda_i L_i), \tag{1}$$

where $Po(k;\beta) = \sum_{i=0}^{k} po(i,\beta)$ is the value of the cumulative distribution function of a Poisson variable with rate β in point k.

The instantaneous service level for the whole system, i.e. the long run proportion of the total demand that can be served directly from stock on hand can be now evaluated by

$$SL^0 = \sum_i SL_i^0 \frac{\lambda_i}{\sum_i \lambda_i}.$$

Service level within response time T. When a demand occurs at a local base i with stock-out, the base has to consider whether it is possible to fulfil the demand by pipeline stock within response time T. This depends on the timing of previous orders issued by the local base. The probability that an arriving customer will be served within time T is calculated in Proposition 1.

Proposition 1. *In a single echelon system, the probability that W_i, the waiting time of an arbitrary customer at base i, is in $(0, T]$ is given by*

$$P(0 < W_i \leq T) = Po(S_i - 1; \lambda_i(L_i - T)) - Po(S_i - 1; \lambda_i L_i).$$

Proof of Proposition 1. Consider the system in equilibrium and look at an arrival at base i. The order triggered by this arrival will satisfy the demand of the S_i- th future customer. The time Z_{S_i} until the S_i- th future customer arrives is Erlang distributed with shape parameter S_i and rate λ_i. Clearly, if the S_i-th future customer has to wait, Z_{S_i} and her waiting time are related by $Z_{S_i} + W_i = L_i$. Thus,

$$\begin{aligned} P(0 < W_i \leq T) &= P(L_i - T \leq Z_{S_i} < L_i) \\ &= [1 - \sum_{n=0}^{S_i-1} \frac{(\lambda_i L_i)^n}{n!} e^{-\lambda_i L_i}] - [1 - \sum_{n=0}^{S_i-1} \frac{[\lambda_i (L_i - T)]^n}{n!} e^{-\lambda_i (L_i - T)}] \\ &= Po(S_i - 1; \lambda_i (L_i - T)) - Po(S_i - 1; \lambda_i L_i). \end{aligned}$$

□

A more general result for an (s, S) inventory system has been proven differently in Kruse (1981) using a compound renewal process for the cumulative demand.

Based on Proposition 1, the proportion of demand at base i satisfied within response time T, SL_i^T, is defined as

$$SL_i^T = SL_i^0 + P(0 < W_i \leq T) = Po(S_i - 1; \lambda_i (L_i - T)).$$

The intuition behind this formula is that if customers accept a response time of T, more demand can be fulfilled because pipeline stocks may arrive within T. This is equivalent to reducing the lead time from L_i to $L_i - T$. The service level within response time T for the whole system, i.e., the long run proportion of the total demand satisfied within time T, can now be calculated by

$$SL^T = \sum_i SL_i^T \frac{\lambda_i}{\sum_i \lambda_i}.$$

Given the distribution of the inventory level at local base i, the long-run average inventory on hand at local base i (EOH_i), i.e., the long-run average physical stock on hand is given by

$$EOH_i = \sum_{n=0}^{S_i-1} (S_i - n) po(n; \lambda_i L_i).$$

The long-run average pipeline stock at base i is given by

$$EPS_i = \lambda_i L_i.$$

5. Evaluation of performance with lateral transshipments

When lateral transshipments are allowed, a base not only fulfils the demand of its own customers but may also need to fulfil lateral transshipment requests from other bases. In this case, the instantaneous service level SL_i^0, the fraction ω_i of customers who face a stock out and are served from the pipeline within time T, the fractions α_{ij} of direct demand at i served by other bases $j \in B$ via lateral transshipments and the fraction θ_i of customers served by i after time T, depend on the corresponding proportions of demands at other bases. We tackle this issue by using an iterative procedure to update the values of SL_i^0, ω_i and α_{ij}, similar to the one proposed by Axsäter (1990), Alfredsson and Verrijdt (1999) and Kutanoglu and Mahajan (2009). Fast convergence of this approach has been reported in the literature for similar models.

We start this section with the description of a queuing model which allows us to approximate the instantaneous service level SL_i^0, ω_i and α_{ij} . We conclude by incorporating all these quantities in an iterative procedure.

Associated $M(n)/D/\infty$ queuing model: Our analysis of each base i relies on approximating the inventory model at base i by a queuing model with an infinite number of servers and state dependent arrival rates. In this model, arrivals correspond to placement of replenishment orders and service time corresponds to the lead time. The number of outstanding orders in our inventory model corresponds to the number of busy servers N_i in the queuing model. We assume that the proportion α_{ij} of direct demand at i satisfied by stock on hand by $j \in B_i$ is known for all the bases $i \in B$.

The rates in the associated queuing model are defined as follows. As long as $N_i < S_i$ (corresponding to the situation with items on stock), demand arrivals are assumed to occur according to a Poisson process with rate

$$\delta_i = \lambda_i + \frac{1}{SL_i^0} \sum_{j|i \in B_j} \alpha_{ji}\lambda_j. \tag{2}$$

In other words, when there is stock on hand at i, the demand rate at i is increased by the proportion of demand from other bases that i will satisfy from its own stock on hand.

When there are $N_i \geq S_i$ busy servers (corresponding to a stock out situation), arrivals are assumed to occur according to a Poisson process with rate

$$\gamma_i = \lambda_i(1 - \frac{1}{1 - SL_i^0} \sum_{j \in B_i} \alpha_{ij}). \tag{3}$$

Thus, when there is no stock on hand, the rate λ_i is reduced by the proportion of direct demand at i that is served by lateral transshipments from stock on hand by bases in B_i.

Observe that in case of identical bases, with lateral transshipment time between any two bases smaller than T, the proportions SL_i^0 and α_{ij} are the same for all bases and $\delta_i SL_i^0 + \gamma_i(1 - SL_i^0) = \lambda_i$, for each $i \in B$. Moreover, if the bases are not identical but $B_i = B$, since $\sum_{i \in B} \sum_{j \in B} \alpha_{ji} = \sum_{i \in B} \sum_{j \in B} \alpha_{ij}$, we have that $\sum_{i \in B} \delta_i SL_i^0 + \gamma_i(1 - SL_i^0) = \sum_{i \in B} \lambda_i$.

Following Dekker et al. (2002) and Lee and Spitler (2006), the steady state distribution of the number of busy servers N_i in this system is given by

$$P(N_i = n) = p_n = \begin{cases} p_0 \frac{(L_i \delta_i)^n}{n!}, & \text{if } n < S_i \\ p_0 \left(\frac{\delta_i}{\gamma_i}\right)^{S_i} \frac{(L_i \gamma_i)^n}{n!}, & \text{if } n \geq S_i \end{cases}$$

or equivalently,

$$P(N_i = n) = \begin{cases} p_0 e^{\delta_i L_i} po(n; \delta_i L_i), & \text{if } n < S_i \\ p_0 e^{\gamma_i L_i} \left(\frac{\delta_i}{\gamma_i}\right)^{S_i} po(n; \gamma_i L_i), & \text{if } n \geq S_i. \end{cases} \tag{4}$$

In (4), p_0 can be now easily calculated based on the normalization constraint $\sum_{n=0}^{\infty} p_n = 1$ and is equal to

$$p_0 = \frac{1}{\sum_{n=0}^{S_i - 1} \frac{(\delta_i L_i)^n}{n!} + (\frac{\delta_i}{\gamma_i})^{S_i} \sum_{n=S_i}^{\infty} \frac{(\gamma_i L_i)^n}{n!}}. \tag{5}$$

As in the case without lateral transshipments, we are interested in the waiting time distribution of an arbitrary arriving customer.

Proposition 2. *In a single echelon system with a base stock policy and arrival rate δ_i, when there are items in stock and γ_i when there is a stock out, the probability that the waiting time W_i of an arbitrary customer in $(0, T]$ is given by*

$$P(0 < W_i \leq T) = p_0 e^{\gamma_i L_i} \left(\frac{\delta_i}{\gamma_i}\right)^{S_i} [Po(S_i - 1; \gamma_i(L_i - T)) - Po(S_i - 1; \gamma_i L_i)],$$

where S_i is the base stock level and p_0 is calculated according to (5).

Proof of Proposition 2. Tag an arbitrary customer who meets a stock out and let W_i be her waiting time. Suppose that just after the arrival of the tagged customer, there are $n \geq S_i + 1$ orders in pipeline, including the order placed at the arrival of the tagged customer. Clearly, there are $n - S_i$ customers waiting. The tagged customer will receive service within T if there will be at least $n - S_i$ items arriving to stock within T.

In Brumelle (1978) it is proven that, conditioned on the fact that there are n busy servers in an $M(n)/G(n)/\infty$ queue, the amounts of work remaining for the n customers in service have the same distribution as n independent variables with equilibrium distribution $H^*(x) = \frac{1}{E(B)} \int_0^x P(B > t)dt$, where B is the work needed to process a customer. In our case, $B = L_i$ and $H^*(x) = \frac{x}{L_i}$ if $x \leq L_i$ and 0 otherwise.

Since for given n, the remaining service time distribution is independent of the arrival rate, we can conclude that, the conditional distribution of the remaining work for the n present customers is the same as the conditional distribution in an $M/G/\infty$ queue with arrival rate γ_i.

Remember that the number of items in pipeline for an inventory model with back orders and constant arrival rate can be analysed with the help of an $M/G/\infty$ queue. Let $\tilde{p_n}$ be the steady state probability that there are n items in pipeline when we set $\delta_i = \gamma_i$ and let W_i' be the waiting time of an arbitrary customer arriving to this system when there are n items in pipeline. Based on the above discussion,

$$P(0 < W_i \leq T |\text{ n in pipeline }) = P(0 < W_i' \leq T |\text{ n in pipeline }),$$

or equivalently,

$$P(0 < W_i \leq T, \text{ n in pipeline }) = P(0 < W_i' \leq T |\text{ n in pipeline })p_n,$$

By Palm's theorem, $\tilde{p_n} = po(n; \gamma_i L_i)$. From (4) it follows that for $n \geq S_i$,

$$p_n = p_0 e^{\gamma_i L_i} \left(\frac{\delta_i}{\gamma_i} \right)^{S_i} \tilde{p_n}.$$

This implies

$$\begin{aligned} P(0 < W_i \le T) &= \sum_{n=S_i}^{\infty} P(0 < W_i \le T, \text{ n in pipeline }) \\ &= \sum_{n=S_i}^{\infty} P(0 < W_i' \le T | \text{ n in pipeline }) p_n \\ &= p_0 e^{\gamma_i L_i} \left(\frac{\delta_i}{\gamma_i}\right)^{S_i} \sum_{n=S_i}^{\infty} P(0 < W_i' \le T | \text{ n in pipeline}) \tilde{p_n}. \end{aligned}$$

Since

$$P(0 < W_i' \le T) = \sum_{n=S_i}^{\infty} P(0 < W_i' \le T | \text{ n in pipeline }) \tilde{p_n},$$

by applying Proposition 1 we obtain

$$P(0 < W_i \le T) = p_0 e^{\gamma_i L_i} \left(\frac{\delta_i}{\gamma_i}\right)^{S_i} [Po(S_i - 1; \gamma_i(L_i - T)) - Po(S_i - 1; \gamma_i L_i)].$$

□

Remark Note that when $\delta_i = \gamma_i$, Proposition 2 degenerates to Proposition 1.

The service levels SL^0 and SL^T will be calculated with the help of the following iterative procedure.

Iterative procedure. Step 1 We start the procedure by setting α_{ij} equal to zero for all $i \in B$ and $j \in B_i$. We calculate the initial SL_i^0 by using (1).

Step 2 Until convergence of the proportions α_{ij} is achieved, repeat steps (2a)-(2d) described below.

2a. Calculate the arrival rate for the associated $M(n)/D/\infty$ queuing models for each base i. Calculate δ_i, respectively γ_i by using (2) and (3).

2b. Calculate SL_i^0 for each base i. Apply (4) for the associated $M(n)/D/\infty$ queuing model with the new rates to find the instantaneous service level SL_i^0:

$$\begin{aligned} SL_i^0 &= P(N_i < S_i) \\ &= p_0 e^{\delta_i L_i} \sum_{n=0}^{S_i - 1} po(n; \delta_i L_i) \end{aligned}$$

$$= p_0 e^{\delta_i L_i} Po(S_i - 1; \delta_i L_i),$$

where p_0 is given by (5).

2c. Calculate ω_i for each base i. By using Proposition 2 and assuming that the PASTA property holds, the fraction ω_i of direct demand of base i that finds a stock out at arrival but can be served by pipeline stock within time T can be estimated by

$$\begin{aligned}\omega_i &= P(0 < W_i \leq T)\\ &= p_0 e^{\gamma_i L_i} (\frac{\delta_i}{\gamma_i})^{S_i} [Po(S_i - 1; \gamma_i(L_i - T)) - Po(S_i - 1, \gamma_i L_i)].\end{aligned}$$

2d. Calculate α_{ij} for each base i and $j \in B_i$. The proportion of direct demand at i that cannot be served within time T by stock on hand or from pipeline stock at i is equal to $1 - SL_i^0 - \omega_i$. Assume that $B_i = \{j_1, ..., j_k\}$. Recall that for direct demand at i that cannot be served by stock on hand or pipeline stock within time T, one checks the stock on hand at bases in B_i, in the prescribed order. The first base with stock on hand, if there is one, will satisfy the demand. Thus, for each $j_l \in B_i$, α_{ij_l} can be estimated by

$$\alpha_{ij_l} = (1 - SL_i^0 - \omega_i)(1 - SL_{j_1}^0)...(1 - SL_{j_{l-1}}^0) SL_{j_l}^0.$$

Note that in this step we have assumed that all bases are independent and the requests that are not satisfied by a base follow a Poisson process.

We denote by δ^*, γ^*, SL_i^{0*}, ω_i^*, the converged values of δ, γ, SL_i^0, and ω_i. For the whole system, the instantaneous service level with lateral transshipments can now be estimated by

$$SL^{0*} = \sum_i SL_i^{0*} \frac{\lambda_i}{\sum_i \lambda_i}$$

The service level within response time T at base i is approximated by

$$SL_i^{T*} = SL_i^{0*} + \omega_i^* + \sum_{j|j \in B_i} \alpha_{ij}^*,$$

while the service level within response time T for the whole system is approximated by

$$SL^{T*} = \sum_i SL_i^{T*} \frac{\lambda_i}{\sum_i \lambda_i}.$$

Furthermore, we can calculate the long-run average inventory on hand at base i by evaluating the expected value of $(S_i - N_i)^+$ from the approximating $M(n)/D/\infty$ system, with steady state probabilities denoted by p_n^*.

$$EOH_i^* = E(IL_i^+) = E[(S_i - N_i)^+] = \sum_{n=0}^{S_i-1} (S_i - n)p_n^*,$$

where p_n^* is given by (4) for rates δ^* and γ^*.

The long-run average pipeline stock at base i now becomes

$$EPS_i^* = \sum_{n=0}^{\infty} np_n^*. \tag{6}$$

Finally, the long-run average demand at base i fulfilled by lateral transshipments from base j can be estimated by

$$ELT_{ij}^* = \alpha_{ij}\lambda_i. \tag{7}$$

Remark It is fairly easy to extend our model to accommodate the relaxed condition of holding back level q_i, where $q_i > 0$ represents the minimum amount of inventory that base i needs to hold when deciding whether to transship its stocks on hand to other bases. We only need to specify the state in the $M(n)/D/\infty$ queue where the arrival rate changes. Particularly, we will have

$$\begin{cases} \delta_i = \lambda_i + \frac{1}{P(N_i < S_i - q_i)} \sum_{j|i \in B_j} \alpha_{ji}\lambda_j, & \text{if } N_i < S_i - q_i \\ \gamma_i = \lambda_i(1 - \frac{\sum_{j \in B_i} \alpha_{ij}}{P(N_i \geq S_i - q_i)}), & \text{if } N_i \geq S_i - q_i. \end{cases}$$

The quality of the iterative procedure will be studied in Section 7.

6. Base stock optimization

In this section, we optimize the base stocks at the local bases such that the total costs of the system are minimized and the system service levels (instantaneous service level and service level within time T) are above given targets.

The total cost TC per time unit comprises the following costs: holding costs for the stocks on hand at all local bases, carrying costs for the stocks in the pipeline between the production plant and all local bases, and lateral transshipment costs for the stocks transshipped between all local bases. In general, the holding cost rate hc includes storage cost, opportunity cost of capital tied up in stocks, insurance cost and costs associated with risk of deterioration or obsolescence; the corresponding pipeline stock cost rate pc includes the same cost elements as the holding costs except for the storage costs; the costs associated with a lateral transshipment requested of base i satisfied by base j, lc_{ij} include the transportation costs from j to i, the insurance costs, the holding costs and the costs associated with risk of deterioration. Since the inventory system is owned by a single principal and controlled centrally, the cost allocation among local bases doe not influence the aggregate profit, being a purely internal transfer price, see Rudi et al. (2001) .

The total cost TC can thus be calculated by

$$TC(\mathbf{S}) = hc \cdot \sum_i EOH_i^* + pc \cdot \sum_i EPS_i^* + \sum_i \sum_{j|j \in B_i} lc_{ij} ELT_{ij}^*,$$

where $\mathbf{S} = (S_1, S_2, .., S_N)$ and EPS_i^*, respectively ELT_{ij}^* are calculated by (6) and (7).

Note that the expected number of units in the pipeline is affected by lateral transshipments since the demand rate at each base depends on the lateral transshipment requests. Therefore, we include the expected pipeline costs in the cost function.

The optimization problem of minimizing the expected costs subject to the customer oriented service level constraints can be formulated as

$$Minimize_{\mathbf{S}} \quad TC(\mathbf{S})$$

$$\text{subject to } SL^{0*}(\mathbf{S}) \geq \phi \text{ and } SL^{T*}(\mathbf{S}) \geq \tau.$$

To solve the optimization problem, we use a complete enumeration over all possible stock profiles $\mathbf{S} = (S_1, S_2, ..., S_N)$ between a lower and an upper bound as in Kutanoglu and Mahajan (2009).

We determine the upper bound of the base stock level at base i, S_i^{max}, by finding the base stock level that achieves the target service levels when the demand rate is $\sum_i \lambda_i$ for all i, when no lateral transshipments are allowed.

We determine the lower bound of the total base stock levels as in Kutanoglu and Mahajan (2009), assuming that all base stocks are pooled together and that they can be delivered to customers instantaneously when demands occur at local bases. This leads to a single echelon system with only one base, and with lead time equal to the minimum of the lead times of all the local bases. Solving this system for the base stock level that achieves the target service levels, we can find the lower bound S^{LB} for the total base stocks. Hence, the total base stocks can range from S^{LB} to $\sum_i S_i^{max}$. We then enumerate all possible stock profiles $\mathbf{S}$ where the total base stocks are in the range $[S^{LB}, \sum_i S_i^{max}]$. For each stock profile $\mathbf{S}$, we check whether it satisfies the service level constraints and calculate the corresponding total inventory cost. The solution is the stock profile that satisfies the service level constraints with minimum total inventory cost.

7. Numerical experiments

In this section we study the numerical performance of the iterative procedure proposed in Section 5 and the effects of lateral transshipments and pipeline stock flexibility on our inventory model.

Numerical performance of the iterative method

In order to evaluate the performance of the iterative method proposed in Section 5, we conducted 6 numerical experiments with identical bases, and 6 numerical experiments with non-identical bases with different demand rate at each base. The experimental inputs are chosen similarly to the experiments in Alfredsson and Verrijdt (1999). Even though service parts are usually slow moving critical parts that require high service levels, for the completeness of the analysis, we also include in our study cases with high demand rates and low service levels that are less likely to appear in practice.

In all our experiments, we consider 3 bases and the time is expressed in days. The local replenishment times are chosen equal to 3 days. The response time is $T = 0.6$ days. The lateral transshipment times between any two bases are 0.5 days, so all lateral transshipments can be delivered within response time. The priority list for each base is $B_1 = \{2,3\}$, $B_2 = \{3,1\}$ and $B_3 = \{1,2\}$. We evaluated the quality of the parameters SL_i^0, ω_i and SL_i^T obtained via the iterative procedure by comparing them with the values obtained via simulation. The results obtained via simulation are average values over 100 runs, each run containing 3650 days.

In the 6 experiments with identical local bases, we considered arrival rates of $0.08, 0.1$ and 0.2. For each arrival rate, the base stock level had values 1 or 2. The results for SL_i^0, ω_i and SL_i^T in the case of identical bases are presented in Table 1. In all our experiments the results obtained via the iterative procedure are close to the ones obtained via simulation. The absolute average error of the instantaneous service level is below 0.01, and of the service level within response time below 0.06. In all the cases, the iterative procedure slightly overestimated the service level within response time. The largest errors occur in Case 5, which is characterized by high demand rates and low base stocks, hence low service levels. Note that since $\sum_{j\in B_i}\alpha_{ij} = SL_i^T - SL_i^0 - \omega_i$ and $\theta_i = 1 - SL_i^T$, the results in Table 1 imply that the iterative procedure also gets values close to the simulation for $\sum_{j\in B_i}\alpha_{ij}$ and θ_i. The average number of iterations required to converge in the cases with identical local bases is 5 with a minimum of 3 iterations and a maximum of 8 iterations.

Table 1 Performance of the iterative procedure for 3 identical local bases

Inputs			SL_i^0		ω_i		SL_i^T	
Cases	λ_i	S_i	Approx[a]	Δ [b]	Approx[a]	Δ[b]	Approx[a]	Δ[b]
1	0.08	1	0.77	0.00	0.05	0.00	0.99	0.02
2		2	0.98	0.00	0.01	0.00	1.00	0.00
3	0.10	1	0.71	0.00	0.06	0.00	0.98	0.03
4		2	0.96	0.00	0.01	0.00	1.00	0.00
5	0.20	1	0.47	-0.01	0.10	0.00	0.88	0.06
6		2	0.88	0.00	0.04	0.00	1.00	0.00

[a] values obtained by the iterative procedure; [b] Δ= values obtained via iterative procedure − values obtained by simulation.

In the 6 experiments with nonidentical bases, all the inputs are the same as in Table 1 with exception of the demand rates. The demand rate at base I is 50% of that at base II and the demand rate at base III is 150% of that at base II. The demand rates at base II were chosen equal to $0.08, 0.1$ and 0.2. The results are presented in Table 2 .

Our approximation is again close to the simulated system performance. The absolute average error of the instantaneous service level is below 0.02 and of the service level within response time below 0.06. Again, the iterative procedure slightly overestimated the service level within response time in all our experiments. The largest errors occur in Case 5a, which is characterized by high

Table 2 Performance of the iterative procedure for 3 non-identical local bases

Inputs				SL_i^0		ω_i		SL_i^T	
Cases	Base	λ_i	S_i	Approx[a]	Δ [b]	Approx[a]	Δ[b]	Approx[a]	Δ[b]
1a	I	0.04	1	0.82	0.00	0.04	0.00	0.99	0.02
	II	0.08	1	0.78	0.00	0.04	0.00	0.99	0.02
	III	0.12	1	0.70	0.00	0.06	0.00	0.99	0.02
2a	I	0.04	2	0.99	0.00	0.00	0.00	1.00	0.00
	II	0.08	2	0.98	0.00	0.01	0.00	1.00	0.00
	III	0.12	2	0.95	0.00	0.02	0.00	1.00	0.00
3a	I	0.05	1	0.76	-0.01	0.05	0.00	0.98	0.03
	II	0.10	1	0.73	0.00	0.05	0.00	0.98	0.03
	III	0.15	1	0.64	0.01	0.07	0.00	0.98	0.02
4a	I	0.05	2	0.99	0.00	0.00	0.00	1.00	0.00
	II	0.10	2	0.97	0.00	0.01	0.00	1.00	0.00
	III	0.15	2	0.93	0.00	0.02	0.00	1.00	0.00
5a	I	0.10	1	0.51	-0.02	0.09	0.01	0.88	0.06
	II	0.20	1	0.49	-0.01	0.09	0.00	0.88	0.06
	III	0.30	1	0.41	0.01	0.10	0.00	0.88	0.06
6a	I	0.10	2	0.94	0.01	0.02	0.00	1.00	0.01
	II	0.20	2	0.89	0.01	0.04	0.00	1.00	0.00
	III	0.30	2	0.80	0.01	0.07	0.00	1.00	0.00

[a] values obtained by the iterative procedure; [b] $\Delta =$ values obtained via iterative procedure $-$ values obtained by simulation.

demand rate and low base stock and hence low service level. The average number of iterations required to converge in the Cases 1a to 6a, is 5.13 with a minimum of 4 iterations and a maximum of 8 iterations.

Based on the results in Table 1 and 2 we conclude that our approximation is quite accurate. The small errors we observe are probably mainly due to the assumption of independent Poisson processes for the lateral transshipment requests between bases, while in fact they are correlated processes.

Benefits of lateral transshipments and pipeline stock flexibility

In order to assess the benefits of using lateral transshipment and pipeline stock flexibility, we consider Cases 1 to 6 again. We compare the performance when both pipeline stock and lateral transshipment flexibility are used to the performance when only one of these sources of flexibility is present. For the cases with lateral transshipments, we use the results obtained by simulation instead of the approximate results, in order to assure a fair comparison to the exact solution of the model without lateral transshipments.

In Table 3 we compare the results for the cases with both pipeline stock and lateral transshipment flexibility included to the results when only the pipeline stock flexibility is present. In this case, since no lateral transshipments are allowed, $\alpha_{ij} = 0$ for all $i, j \in B$.

Table 3 Benefits of lateral transshipment flexibility for 3 identical bases

Inputs			Pipeline and Lateral Transship.				Only Pipeline			Δ[a]			
Cases	λ_i	S_i	SL_i^0	ω_i	α_i[b]	SL_i^T[c]	SL_i^0	ω_i	SL_i^T[c]	SL_i^0	ω_i	α_i[b]	SL_i^T
1	0.08	1	0.77	0.05	0.16	0.97	0.79	0.04	0.83	-0.02	0.01	0.16	0.15
2		2	0.98	0.01	0.02	1.00	0.98	0.01	0.98	0.00	0.00	0.02	0.02
3	0.10	1	0.71	0.06	0.19	0.96	0.74	0.05	0.79	-0.03	0.01	0.19	0.17
4		2	0.96	0.01	0.02	1.00	0.96	0.01	0.98	0.00	0.00	0.02	0.02
5	0.20	1	0.48	0.09	0.25	0.82	0.55	0.07	0.62	-0.07	0.02	0.25	0.20
6		2	0.87	0.04	0.08	0.99	0.88	0.04	0.92	0.00	0.01	0.08	0.08

[a] $\Delta =$ values for the system with pipeline and lateral transshipment flexibility $-$ values for the system with pipeline flexibility, but no lateral transshipments; [b] $\alpha_i = \sum_{j \in B_i} \alpha_{ij}$ represents the proportion of direct demand at base i satisfied by lateral transshipments;
[c] results obtained by simulation.

The results show that the service level within response time benefits most from the presence of lateral transshipments, because it enables local bases to share resources. This benefit is more pronounced when the demand rate is high and the base stock level is low. The higher the direct demand rate at a base, the higher the proportion of demand that is satisfied by lateral transshipments. On the other hand, in almost all cases, lateral transshipment flexibility does not improve the instantaneous service level, because the stocks being transshipped cannot be utilized immediately due to the non-negligible transshipment time.

Next we study the effect of lateral transshipments for the cases with non-identical bases. In Table 4 we compare for Cases 1a to 6a the results when both pipeline stock and lateral transshipment flexibility are used to the results when only the pipeline stock flexibility is present.

Table 4 shows that lateral transshipment flexibility improves the service levels within response time SL_i^T also for nonidentical bases. The reason is that more demands can be fulfilled via lateral transshipments, as indicated by α_i. These effects are more pronounced in cases with higher demand rates and lower base stock levels. Among the local bases, the base with highest demand rate benefits most from lateral transshipments. The instantaneous service level SL_i^0 is not affected much but the

Table 4 Benefits of lateral transshipment flexibility for 3 non-identical bases

Inputs				Pipeline and Lateral Trans.				Only Pipeline			Δ [a]			
Cases	Base	λ_i	S_i	SL_i^0	ω_i	α_i[b]	SL_i^T[c]	SL_i^0	ω_i	SL_i^T[c]	SL_i^0	ω_i	α_i[b]	SL_i^T
1a	I	0.04	1	0.82	0.04	0.12	0.97	0.89	0.02	0.91	-0.06	0.01	0.12	0.06
	II	0.08	1	0.78	0.04	0.15	0.97	0.79	0.04	0.83	-0.01	0.00	0.15	0.15
	III	0.12	1	0.70	0.06	0.22	0.97	0.70	0.05	0.75	0.00	0.01	0.22	0.22
2a	I	0.04	2	0.99	0.00	0.01	1.00	0.99	0.00	1.00	0.00	0.00	0.01	0.00
	II	0.08	2	0.98	0.01	0.02	1.00	0.98	0.01	0.98	0.00	0.00	0.02	0.02
	III	0.12	2	0.95	0.02	0.03	1.00	0.95	0.02	0.97	0.00	0.00	0.03	0.03
3a	I	0.05	1	0.77	0.05	0.14	0.95	0.86	0.03	0.89	-0.09	0.02	0.14	0.07
	II	0.10	1	0.72	0.05	0.18	0.96	0.74	0.05	0.79	-0.02	0.01	0.18	0.17
	III	0.15	1	0.64	0.07	0.25	0.96	0.64	0.06	0.70	0.00	0.01	0.25	0.26
4a	I	0.05	2	0.99	0.00	0.01	1.00	0.99	0.00	0.99	0.00	0.00	0.01	0.01
	II	0.10	2	0.96	0.01	0.02	1.00	0.96	0.01	0.98	0.00	0.00	0.02	0.02
	III	0.15	2	0.93	0.02	0.05	1.00	0.92	0.02	0.95	0.00	0.00	0.05	0.05
5a	I	0.10	1	0.53	0.09	0.20	0.82	0.74	0.05	0.79	-0.21	0.04	0.20	0.03
	II	0.20	1	0.49	0.09	0.24	0.82	0.55	0.07	0.62	-0.06	0.02	0.24	0.20
	III	0.30	1	0.40	0.10	0.32	0.82	0.41	0.08	0.49	-0.01	0.02	0.32	0.33
6a	I	0.10	2	0.93	0.02	0.04	0.99	0.96	0.01	0.98	-0.03	0.01	0.04	0.02
	II	0.20	2	0.88	0.04	0.07	0.99	0.88	0.04	0.92	0.00	0.00	0.07	0.08
	III	0.30	2	0.79	0.07	0.13	0.99	0.77	0.06	0.84	0.02	0.01	0.13	0.16

[a] Δ= values for the system with pipeline and lateral transshipment flexibility − values for the system with pipeline flexibility, but no lateral transshipments; [b] $\alpha_i = \sum_{j \in B_i} \alpha_{ij}$ represents the proportion of direct demand at base i satisfied by lateral transshipments;
[c] results obtained by simulation.

service level within response time SL_i^T is improved dramatically due to the help from other bases. On the other hand, the bases with lower demand rates suffer from lower instantaneous service levels because they have to share their resources.

In order to assess the effect of pipeline stock flexibility, we compared the simulated results for Cases 1 to 6 to simulated results with the same inputs but excluding the pipeline stock flexibility. The comparison is presented in Table 5.

Table 5 Benefits of pipeline stock flexibility for 3 identical bases

Inputs			Pipeline and Lateral Trans.				Only Lateral Trans.				Δ^a			
Cases	λ_i	S_i	SL_i^0	ω_i	α_i[b]	SL_i^T[c]	SL_i^0	ω_i	α_i[b]	SL_i^T[c]	SL_i^0	ω_i	α_i[b]	SL_i^T
1	0.08	1	0.77	0.05	0.16	0.97	0.77	0.00	0.20	0.97	0.00	0.05	-0.03	0.01
2		2	0.98	0.01	0.02	1.00	0.98	0.00	0.02	1.00	0.00	0.01	-0.01	0.00
3	0.10	1	0.71	0.06	0.19	0.96	0.71	0.00	0.23	0.94	0.00	0.06	-0.05	0.01
4		2	0.96	0.01	0.02	1.00	0.96	0.00	0.04	1.00	0.00	0.01	-0.01	0.00
5	0.20	1	0.48	0.09	0.25	0.82	0.47	0.00	0.31	0.78	0.00	0.09	-0.06	0.04
6		2	0.87	0.04	0.08	0.99	0.88	0.00	0.12	0.99	0.00	0.04	-0.04	0.00

[a] Δ= values for the system with pipeline and lateral transshipment flexibility − values for the system with pipeline flexibility, but no lateral transshipments; [b] $\alpha_i = \sum_{j \in B_i} \alpha_{ij}$ represents the proportion of direct demand at base i satisfied by lateral transshipments;
[c] results obtained by simulation.

As expected, the results show that although pipeline stock flexibility has little impact on the instantaneous service levels SL_i^0, it improves the service levels within response time SL_i^T. This benefit is more pronounced when the demand rate is high and the base stock level is low. Furthermore, with pipeline stock flexibility we avoid unnecessary lateral transshipments, indicated by lower values for α_i. This aspect of pipeline stock flexibility is especially obvious in cases with high demand rates and low base stocks.

Effects of lateral transshipment and pipeline stock flexibility on the optimal stock levels

To study the effect of the pipeline stock and lateral transshipment flexibility on costs, we consider again 6 cases with 3 bases (see Table 6). The arrival rates at bases, the leadtimes, the time limit T and the priority lists for lateral transshipments are as for Cases 1 to 6. The pipeline stock cost rate pc is set to be €24 per unit per day. The lateral transshipment cost lc_{ij} is set to be €500 per shipment. In experiments 2b, 4b and 6b, the holding cost rate at all bases is equal to €30 per unit per day. In the other cases, the holding cost rate at base III is equal to €60 per unit per day. The service level requirements are $\phi = 0.90$ and $\tau = 0.98$. The results are obtained via the approximation procedure described in Section 5.

In order to assess the effect of the lateral transshipment flexibility on the costs, we compared the results of two sets of experiments: in the first one, both pipeline stock and lateral transshipment flexibility are used, while in the second only the first flexibility is used. In both cases, the base stock levels are optimized with respect to the service level constraints. The optimal base stock levels for the two set-ups and the associated costs are presented in the second and third group of columns of Table 6. Finally, the savings gained by using the lateral transshipment flexibility are presented in the last column.

The results indicate that lateral transshipments are more beneficial when the demand rate is high and the holding cost rate is high compared to the lateral transshipment cost. If no lateral transshipments are possible, more items are kept in stock, resulting in higher holding costs. When the holding cost rate at base III is equal to 60 and lateral transshipments are allowed, the base

Table 6 Benefits of lateral transshipment flexibility for 3 bases

Inputs				Pipeline Stock and Lateral Trans. Flexibility					Only Pipeline Stock Flexibility					Savings
Cases	Base	λ_i	hc_i	S_i^*	HC^a	PC^b	LC^c	TC ^{d}	S_i^*	HC^a	PC^b	LC^c	TC ^{d}	ΔTC^e
1b	I	0,08	30	1	24,00	4,91	6,42	153,70	2	52,86	5,76	0	175,86	22,16
	II	0,08	30	2	51,82	6,56	0,78		2	52,86	5,76	0		
	III	0,08	30	2	52,76	5,81	0,63		2	52,86	5,76	0		
2b	I	0,08	30	2	51,82	6,56	0,78	177,71	2	52,86	5,76	0	228,73	51,02
	II	0,08	30	2	52,76	5,81	0,63		2	52,86	5,76	0		
	III	0,08	**60**	1	48,01	4,91	6,42		2	105,72	5,76	0		
3b	I	0,10	30	2	51,04	7,20	1,17	178,22	2	51,12	7,20	0	204,84	26,62
	II	0,10	30	2	51,04	7,20	1,17		2	51,12	7,20	0		
	III	0,10	30	2	51,04	7,20	1,17		3	81,01	7,20	0		
4b	I	0,10	30	3	79,25	8,60	0,15	208,59	2	51,12	7,20	0	255,96	47,37
	II	0,10	30	2	51,21	7,07	1,13		3	81,01	7,20	0		
	III	0,10	**60**	1	45,59	5,93	9,66		2	102,23	7,20	0		
5b	I	0,20	30	2	43,28	13,56	7,22	216,16	3	72,11	14,40	0	259,54	43,38
	II	0,20	30	2	42,49	14,21	7,82		3	72,11	14,40	0		
	III	0,20	30	3	70,78	15,42	1,37		3	72,11	14,40	0		
6b	I	0,20	30	3	70,78	15,42	1,37	258,65	3	72,11	14,40	0	331,66	73,01
	II	0,20	30	2	43,28	13,56	7,22		3	72,11	14,40	0		
	III	0,20	**60**	2	84,98	14,21	7,82		3	144,23	14,40	0		

[a] total holding costs; [b] total pipeline costs; [c] total lateral transshipment costs; [d] total costs; [e] ΔTC= total costs with only lateral transshipment flexibility − total costs with both pipeline stock and lateral transshipment flexibility.

stock level at base III decreases, while the base stock level at base I increases (see Cases 1 and 2, second group of columns). Due to the low holding costs at I, it is cheaper to reach the overall desired service levels by lateral transshipments between bases I and III than by holding stock at base III. Note that base III requests lateral transshipments first from base I, then from base II. When lateral transshipments are not allowed, the overall service level is achieved by increasing the service level at one of the bases with low holding costs, thus sacrificing the service level at a base with high holding costs (see Cases 3 and 4, third group of columns). The pipeline costs are not affected by lateral transshipment flexibility, since they represent holding costs from the supplier to a base, and are incurred for all items in the same way.

We assess the effect of pipeline stock flexibility by comparing the results of the base stock optimization when both pipeline stock and lateral transshipment flexibility are considered with the results when only lateral transshipment flexibility is used. The optimal base stock levels for both situations and the associated costs are presented in Table 7.

The results show that the benefits of pipeline stock flexibility are larger when the demand rates are higher. Our experiments indicate that including pipeline stock flexibility leads to lower lateral transshipment costs, since unnecessary transshipments are avoided. In our settings, the pipeline stock information did not influence the optimal base stock levels, thus almost the same holding costs were incurred with or without the pipeline stock flexibility. When the holding cost rates differ

Table 7 Benefits of pipeline stock flexibility for 3 bases

Inputs				Pipeline Stock and Lateral Trans. Flexibility					Only Lateral Trans. Flexibility					Savings
Cases	Base	λ_i	hc_i	S_i^*	HC^a	PC^b	LC^c	TC [d]	S_i^*	HC^a	PC^b	LC^c	TC [d]	ΔTC^e
1b	I	0,08	30	1	24,00	4,91	6,42	153,70	1	24,08	4,74	7,89	155,78	2,08
	II	0,08	30	2	51,82	6,56	0,78		2	51,61	6,71	1,25		
	III	0,08	30	2	52,76	5,81	0,63		2	52,71	5,83	0,96		
2b	I	0,08	30	2	51,82	6,56	0,78	177,71	2	51,61	6,71	1,25	179,86	2,15
	II	0,08	30	2	52,76	5,81	0,63		2	52,71	5,83	0,96		
	III	0,08	**60**	1	48,01	4,91	6,42		1	48,15	4,74	7,89		
3b	I	0,10	30	2	51,04	7,20	1,17	178,22	2	51,00	7,20	1,78	179,93	1,71
	II	0,10	30	2	51,04	7,20	1,17		2	51,00	7,20	1,78		
	III	0,10	30	2	51,04	7,20	1,17		2	51,00	7,20	1,78		
4b	I	0,10	30	3	79,25	8,60	0,15	208,59	3	78,86	8,91	0,30	211,32	2,73
	II	0,10	30	2	51,21	7,07	1,13		2	51,24	7,01	1,69		
	III	0,10	**60**	1	45,59	5,93	9,66		1	45,79	5,68	11,84		
5b	I	0,20	30	2	43,28	13,56	7,22	216,16	2	43,42	13,27	10,52	224,05	7,89
	II	0,20	30	2	42,49	14,21	7,82		2	42,39	14,09	11,70		
	III	0,20	30	3	70,78	15,42	1,37		3	70,20	15,84	2,62		
6b	I	0,20	30	3	70,78	15,42	1,37	258,65	3	70,20	15,84	2,62	266,43	7,78
	II	0,20	30	2	43,28	13,56	7,22		2	43,42	13,27	10,52		
	III	0,20	**60**	2	84,98	14,21	7,82		2	84,77	14,09	11,70		

[a] total holding costs; [b] total pipeline costs; [c] total lateral transshipment costs; [d] total costs; [e] ΔTC= total costs with only lateral transshipment flexibility − total costs with both pipeline stock and lateral transshipment flexibility.

among the bases, the same phenomena as in Table 6 can be observed.

8. Case study

We have applied our models in a case study for a manufacturer in the dredging industry, which builds dredging vessels and supplies equipment and control systems to customers worldwide. Among the assortment of all service parts, we selected the most important one for demonstration in this paper, namely an impeller.

The impeller is a critical component (usually made of cast iron) of a centrifugal pump in dredgers, which accelerates a combination of water and several soils, such as sand, silt and gravel through the piping system. Usually, the impeller is worn out faster than the pump casing and it has to be replaced with a new one to keep the pump running. Moreover, the impeller weighs approximately 1700 kg (depending on the size), meaning that it is way too costly to transport it by air. Consequently, slow sea transport is needed, which takes much more time, leading to more pipeline stocks and non-negligible lateral transshipment times.

The company has a two-echelon inventory system with a central depot in the Netherlands, which repairs all the broken impellers. The time required to repair an impeller is typically around $L_0 = 35$ weeks. There are 3 operating bases, located in Shanghai (Base 1), Singapore (Base 2), and Dubai (Base 3) respectively. The lead time between the central depot and these bases is $L_1 = 8$ weeks, $L_2 = 7$ weeks, and $L_3 = 6$ weeks respectively.

In case of a stock-out, a base may request a lateral transshipment from other bases based on pre-specified rules, as in Section 3.

The lateral transshipment time is 2 weeks between Base 1 and 2, 3 weeks between Base 2 and 3, and 5 weeks between Base 1 and 3. Moreover, the acceptable response time is 3 weeks. As a result, the priority lists pre-specified by the company are $B_1 = \{2\}$, $B_2 = \{1, 3\}$, and $B_3 = \{2\}$.

Empirical data regarding the impeller was collected and is used as input to our model. The demand rates presented below are realistic but fictitious for confidentiality reasons. The demand rates at the bases are $\lambda_1 = 0.4$ units/week, $\lambda_2 = 0.1$ units/week, and $\lambda_3 = 0.2$ units/week. Thus, the demand rate at the central depot is $\lambda_0 = 0.7$ units/week. The target customer-oriented service levels are 90% probability of instantaneous delivery and 98% probability of delivery within 3 weeks.

The cost parameters for the impeller in this service network are estimated by an industrial expert. The holding cost is around 38 euros per unit per week; the pipeline stock cost is 24 euros per unit per week; the transshipment cost is 1800 euros per unit between Base 1 and 2, 2100 euros per unit between Base 2 and 3, and 2500 euros per unit between Base 1 and 3.

In order to apply our analysis to the case study, we need to extend our model to a two-echelon inventory model. Denote by L_0 the leadtime between plant and central depot. Moreover, the central depot has a limited capacity. Thus, the local replenishment lead time may increase when the central depot has a stock out.

Since local bases are replenished on a one-for-one, first-come first served basis from the central depot, the demand $\{D_0(t)\}$ at the central depot is a superposition of the demands at the local bases. Since the demand processes at the local bases are independent Poisson processes, $D_0(t)$ is also a Poisson process with rate $\lambda_0 = \sum_i \lambda_i$, regardless of whether there are lateral transshipments among the local bases (Lee, 1987).

By Palm's Theorem (Palm, 1938) applied to the central depot, the long-run average inventory on hand at the central depot (EOH_0) is given by

$$EOH_0 = \sum_{n=0}^{S_0-1} (S_0 - n) po(n; \lambda_0 L_0),$$

The long-run average number of back orders at the central depot,i.e., the long-run average number of units that have been requested but not yet delivered, can be calculated by

$$EBO_0 = \lambda_0 L_0 - S_0 + \sum_{n=0}^{S_0-1} (S_0 - n) po(n; \lambda_0 L_0)$$

According to *Little's Law* (Little, 1961), the average delay at the central depot is $E(W_0) = EBO_0/\lambda_0$. This average delay is the same for all local bases because of the Poisson demand at the central depot and the first-come first-served assumption (Axsäter, 2006).

In the two-echelon system, the lead time at a base becomes a stochastic variable depending on the waiting time at the central depot. Because of this dependence, successive lead times at local bases are not independent and therefore Palm's Theorem (Palm, 1938) can only be used as an approximation. By applying this approximation, which is known under the name METRIC-approximation (Sherbrooke, 1968), we can replace the stochastic lead time at base i by $\overline{L_i} = L_i + E(W_0)$. This approach is widely regarded as a reasonable approximation (Axsäter,1990).

Using the METRIC approximation, we can easily adapt Proposition 2 and the iterative procedure in Section 5 for the two-echelon inventory system. Furthermore, we also need to add the inventory holding costs and pipeline stock costs at the central depot to the costs of all local bases for cost evaluation.

Applying the base stock optimization solution described in Section 6 for the two echelon inventory system, we obtain the optimal base stock allocation **S**={24, 8, 3, 4}, i.e., 24 at the central depot in the Netherlands, 8 at the base in Shanghai, 3 base stocks at the base in Singapore and 4 at the base in Dubai. The corresponding minimum cost is 570.22 euros per week. It breaks down to total inventory holding cost per week, total pipeline stock cost per week, and total lateral transshipment cost per week (see Table 8).

Table 8 Total costs per week breakdown

Average cost per week	Total	Holding	Pipeline Stock	Lateral Trans.
Euro	570,22	356,48	179,67	34,06
Percent	100%	62,52%	31,51%	5,97%

As we can see in Table 8, the total holding costs take the largest share of the total cost. However, the total pipeline costs account for 31.51% of the total cost. In fact, the total expected pipeline stocks are more than half of the total expected stock on hand. This is because the lead time of these slow moving items is quite long. For instance, Shanghai has a demand rate of $\lambda_1 = 0.4$ units/week or equivalently, a mean time between consecutive demands of 2.5 weeks, which is much shorter than the lead time of 8 weeks. As a result, we cannot disregard the pipeline stocks in the inventory control decisions.

The achieved instantaneous service level is 90.29%, and the probability of service within 3 weeks is 99.77%, above the 90% and 98% targets. They are not exactly equal to the targets due to integer solutions for the base stocks.

Next, we investigate the economic benefits of lateral transshipment flexibility and pipeline stock flexibility. We compare our results from the case study, with the results from the model without lateral transshipments, and with another model which disregards the pipeline stocks. Since our models approximate the system performance and costs accurately, as shown in Section 7, it is possible to use the approximate results for practical purpose in daily business operations. The results are presented in Table 9.

Table 9 Economic benefits of lateral transshipment and pipeline stock flexibility

Model	S_0^*	S_1^*	S_2^*	S_3^*	TC	Savings
Lateral trans. &pipeline stock	24	8	3	4	571,18	
Only pipeline stock	25	8	3	4	563,17	7,05
Only lateral transshipment	25	8	3	4	657,71	- 87,49

Comparing the results from the model with both pipeline and lateral transshipment flexibility to the models with either pipeline or lateral transshipment flexibility, we conclude that having only lateral transshipment flexibility increases the costs by 87.49 euro, while having only pipeline stock is a little beneficial in this case.

9. Conclusion

This paper assesses the effect of pipeline stock flexibility in one or two-echelon inventory models where the lateral transshipment time is not negligible. We introduce customer-oriented service levels, expressed by the probability of instantaneous service and the probability of service within a certain response time. We propose approximations based on queuing models with state dependent arrival rates. Via extensive numerical experiments, we show that our approximations perform well in terms of both system performance and costs. Our numerical experiments indicate that including both lateral transshipment and pipeline stock flexibility in inventory decisions is more beneficial than including lateral transshipments alone. The magnitude of this effect is higher for high demand rates and high lateral transshipment costs. This conclusion is also supported by our findings in a case study for a market leader in dredging industry. As a result of our research, we recommend the introduction of pipeline stock information such as the track and trace information from freight carriers in existing ERP systems.

Acknowledgments

This research has been made possible with support of Transumo (see www.transumo.nl). We thank the two anonymous referees for their valuable comments and suggestions that have significantly improved the quality of the paper. We also thank Ricardo van Gelder who did preliminary research on this topic. We appreciate the help and insights of Jan Brinkhuis and Willem van Jaarsveld.

References

Alfredsson, P. and Verrijdt, J., 1999. Modeling emergency supply fexibility in a two-echelon inventory system. *Management Science.* **45(10)**, 1416–1431.

Axsäter, S., 1990. Modelling emergency lateral transshipments in inventory systems. *Management Science.* **36(11)**, 1329–1338.

Axsäter, S., 2003. A new Decision Rule for Lateral Transshipments in Inventory Systems. *Management Science.* **49(9)**, 1168-1179.

Axsäter, S., 2006. *Inventory Control*, 2nd edition. International Series in Operations Research & Management Science. New York: Springer Science.

Banerjee, A., Burton, J. and Banerjee, S., 2003. A simulation study of lateral shipments in single supplier, multiple buyers supply chain networks. *International Journal of Production Economics.* **81-82**, 103–114.

Brumelle, S. L., 1978. A generalization of Erlang's loss system to state dependent arrivals and service rates. *Mathematics of Operations Research.***(3)(1)**, 10-16.

Burton, J. and Banerjee, A., 2005. Cost-parametric analysis of lateral transshipment policies in two-echelon supply chains. *International Journal of Production Economics.* **93-94**, 169–178.

Dekker, R., Hill, R.M., Kleijn, M.J., and Teunter, R.H., 2002. On the (s - 1, s) lost sales inventory model with priority demand classes. *Naval Research Logistics.* **49**, 593–610.

Diks, E.B. and De Kok, A.G., 1996. Controlling a divergent 2-echelon network with transshipments using the consistent appropriate share rationing policy. *International Journal of Production Economics.* **45**, 369–379.

Grahovac, J. and Chakravarty, A., 2001. Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items. *Management Science.* **47(4)**, 579–594 .

Howard, C., Reijnen, I.C., Marklund, J. and Tan, T., 2010. Using pipeline information in a multi-echelon spare parts inventory system. BETA working paper, No. 330, Eindhoven University of Technology.

Kruse, K., 1981. Waiting time in a continuous review (s, S) inventory system with constant lead times. *Operations Research.* **29(1)**, 202–207.

Kutanoglu, E. and Mahajan, M., 2009. An inventory sharing and allocation method for a multi-location service parts logistics network with time-based service levels. *European Journal of Operational Research.* **194(3)**, 728–742.

Lee, D.C., Spitler, S.L., 2006. On the $M(n)/G/infty$ steady-state distribution. *Performance Evaluation.* **63(12)**, 1157–1164.

Lee, H., 1987. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science.* **33(10)**, 1302–1316.

Palm, C., 1938. Analysis of the Erlang traffic formula for busy signal arrangements. *Ericsson Technics.* **5**, 39–58.

Paterson, C., Kiesmüller, G., Teunter, R., Glazebrook K., 2011. Inventory models with lateral transshipments: A review. *European Journal of Operational Research.* **210 (2)**, 125–136.

Rudi, N., Kapur, S., and Pyke, D. F., 2001. A two-location inventory model with transshipment and local decision making. *Management Science.* **41(12)**, 1668–1680.

Sherbrooke, C., 1968. METRIC: A multi-echelon technique for recoverable item control. *Operations Research.* **16(1)**, 122–141.

Tagaras, G., Vlachos, D., 2002. Effectiveness of stock transshipment under various demand distributions and nonnegligible transshipment times. *Production and Operations Management.* **11(2)**, 183–198.

Wong, H., Cattrysse, D. and Van Oudheusden, D., 2005. Stocking decisions for repairable spare parts pooling in a multi-hub system. *International Journal of Production Economics.* **93-94**, 309–317.