A new partial pooling structure for spare parts networks

A.A. Kranenburg

Centre for Quantitative Methods CQM B.V.

P.O. Box 414, 5600 AK Eindhoven, The Netherlands

Phone: + 31 40 750 2322, Fax: + 31 40 750 1699, E-mail: kranenburg@cqm.nl

G.J. van Houtum

Department of Technology Management, Technische Universiteit Eindhoven

P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Phone: + 31 40 247 5163, Fax: + 31 40 246 4596, E-mail: g.j.v.houtum@tm.tue.nl

Version date: July 5, 2008

Abstract

Motivated by real-life spare parts networks, we introduce a new spare parts inventory model with lateral transshipment. We consider a multi-item, multi-location, single-echelon system with base stock control and aggregate mean waiting time constraints. The local warehouses are divided into two types: main and regular local warehouses. Lateral transshipment is allowed from main local warehouses only. A practical advantage of this structure is that only a limited number of local warehouses has to be equipped to provide lateral transshipment. This structure represents a new form of partial pooling, with no pooling (zero main locals) and full pooling (zero regular locals) as special cases. We develop an accurate and fast approximate evaluation method, and exploit this method in a heuristic procedure for the base stock level determination. We show that only a small number of main locals is sufficient to obtain most of the full pooling benefits. We also apply our methods to case data of ASML, an original equipment manufacturer in the semiconductor supplier industry, and show that their operations can be significantly improved by our new model.

Keywords: Inventory, spare parts, pooling, lateral transshipment

1 Introduction

We consider the inventory control for networks consisting of multiple local warehouses, where spare parts of expensive capital goods are kept on stock. When in such a network a demand occurs at a local warehouse that is out-of-stock at that moment, one often applies a *lateral transshipment* from another local warehouse. As an illustration, we describe the supply chain of spare parts at ASML, an original equipment manufacturer (OEM) in the semiconductor supplier industry.

Semiconductors are produced in fabs in Europe, North-America, and Asia. Setting up a fab represents a billion dollar investment with a significant portion in equipment. In semiconductor

manufacturing, the most critical production steps are to be made with step&scan systems. ASML is one of the few companies that designs, builds, and supplies this kind of equipment. The opportunity cost in case of lost production when such equipment is down is very high, and potentially could run into millions of euros. Therefore, it is important that the availability of the systems is high, and that down-situations are recovered quickly. Spare parts (service parts) are often needed in the recovery sequence and need to be available quickly. In service level agreements, target aggregate mean waiting times are specified that have to be met by ASML. Targets are set for groups of machines, e.g. for all ASML machines at one fab, or for all ASML machines within that fab that are used for the same functional step or type of chip. Typically, customers require an availability of their systems that corresponds to an aggregate mean waiting time (per demand) of a few hours only. In the fabs themselves, no spare parts are kept on stock, but, typically, an ASML warehouse is in close distance to the customer site. If a customer needs a spare part, it will be delivered from the local warehouse if available there. Stock in the local warehouse is replenished from a central warehouse in The Netherlands, which costs about two weeks time. If a customer needs a spare part that is not available in the local warehouse, it would take too long to wait for a normal replenishment from the central warehouse. Therefore, an emergency replenishment can be carried out from the central warehouse to the local warehouse in that case. For an emergency shipment, at present, up to 48 hours is needed if the local warehouse is in another continent than the central warehouse. This is shorter than the normal replenishment time, but still quite long compared to an aggregate mean waiting time of a few hours. Thus, another option is used preferably, to quickly provide the customer with a spare part upon request: *lateral transshipment*. If a local warehouse that faces a demand is out of stock, neighboring local warehouses, e.g. local warehouses in the same region or continent, are checked for availability of the requested stock-keeping unit (SKU). If the part is available in one of these local warehouses, a lateral transshipment will take place. This costs only about 10 to 20 hours time, and thus drastically reduces the waiting time for the spare part compared to an emergency shipment from the central warehouse. At ASML, in daily execution, lateral transshipment has been applied for years. However, in the *planning phase* in which base stock levels are determined, the option of lateral transshipment was *not* taken into consideration.

The contribution of this paper is as follows:

- (i) *We introduce a network structure that is general and matches with structures observed in practice.* With respect to lateral transshipment, we distinguish two types of local warehouses in our network: *main* and *regular* local warehouses. While both main and regular local warehouses can receive a lateral transshipment, only main local warehouses are allowed to be the supplier of a lateral transshipment. Our network structure is interesting from a theoretical point of view. Consider the two extreme situations: on the one hand a situation with regular local warehouses only, and on the other hand a situation with main local warehouses

only. The former constitutes the single-location situation without any lateral transshipment between local warehouses, or in other words, the *no pooling* situation. The latter case equals the situation with *full pooling*, i.e., where lateral transshipment is allowed between all local warehouses. Thus, *our network structure is a generalization* of which both the no pooling and full pooling situations are special cases. In general, our network structure leads to *partial pooling*, where only some of the local warehouses, namely the main local warehouses, are allowed to provide lateral transshipment. With our general network structure, we are able to compare no, partial and full pooling with respect to their performance and cost.

Our network structure also has a high practical relevance. In fact, the motivation for this network structure comes from the situation that we encountered at ASML: there exist differences between local warehouses. Some local warehouses are larger (have more inventory) since they serve larger or more customers and observe larger demand rates than other local warehouses. Some local warehouses have longer operating hours than others (for example also during the night). Further, some local warehouses are close to airports with many flights (hubs). It is desired that if a lateral transshipment is needed, it will be supplied by a local warehouse that can provide it quickly. In other words, especially local warehouses with the characteristics described above are suitable candidates to be main local warehouses. From the 19 local warehouses of ASML that are located in the United States of America, 4 local warehouses were identified as main local warehouses, for the above reasons. The other 15 local warehouses are regular local warehouses.

Our network structure fits also for networks with so-called quick response stocks, which are being used by several companies; see e.g. Rijk (2007). These companies do not allow lateral transshipment between real local warehouses, but they have installed extra warehouses from where parts can be delivered in case a local warehouse is out of stock. The real local warehouses may be modeled as regular local warehouses in our network structure, and the extra warehouses may be modeled as main local warehouses. In this case, the main local warehouses receive no direct demands from customers and their only function is to be supplier of lateral transshipment.

- (ii) *We develop an approximate evaluation method for instances of real-life size.* The structure of our problem is such that the performance evaluation of the whole system with multiple SKU-s can be decomposed into an evaluation per SKU. Instances with a small number of local warehouses and small base stock levels can be analyzed exactly via a Markov process description. However, this does not work for instances of real-life size. Therefore we develop an approximate evaluation method that is based on the following main idea: (i) We decouple the network into individual local warehouses; (ii) We assume that the extra demand processes

in main local warehouses due to requests for lateral transshipment are Poisson processes. Under the latter assumption, each individual warehouse can be analyzed analytically, and an iterative algorithm is needed to determine the demand rates for lateral transshipment. We show that the resulting approximate evaluation method is accurate and fast.

- (iii) *We show that via partial pooling the major part of the full pooling benefits is obtained.* The approximate evaluation model is used in a greedy heuristic for the multi-item spare parts model with aggregate mean waiting time constraints. Based on this multi-item model, we show that partial pooling performs very well compared to full pooling. If only a few of the local warehouses are allowed to provide lateral transshipment, then already a substantial part of the full pooling benefits is obtained. This result has an important practical implication: only a few, well chosen local warehouses have to be equipped to provide lateral transshipment in order to obtain most of the benefits of full pooling.
- (iv) *Our method leads to substantial cost savings for a case study with ASML data.* Previously, ASML used lateral transshipment between local warehouses in daily operation, but it did not take lateral transshipment into account when planning the base stock levels in the local warehouses. We show for data sets obtained from ASML that substantial savings (in the order of 30%) are realized when lateral transshipment is taken into account in the spare parts inventory planning method. Based on our case study, ASML decided to implement our greedy heuristic for their spare parts inventory model. They use our algorithm since early 2005 and succeeded to reduce both waiting times and cost substantially.

A lot of literature is available on the topic of lateral transshipment. For an overview, see Wong et al. (2006), where lateral transshipment literature is categorized in terms of the number of echelons, the number of items, periodic or continuous review, inventory control policy, and the type of analysis done: exact or approximate evaluation, optimization or approximation. Our work can be described as a single-echelon (but with a special structure within that echelon), multi-item, continuous review model with base stock policies. For this setting, we provide a new method for approximate evaluation that is applicable for large, real-life instances. Furthermore, we provide an approximation using a greedy-type algorithm.

Wong et al. (2006) itself describe a continuous-review, multi-item spare parts system with two local warehouses with lateral transshipment and base stock control. Exact analysis of policies is done using a Markov process description. If the warehouse that faces a demand is out of stock, lateral transshipment from the other local warehouse is applied. If the other local warehouse has no stock on hand, an emergency replenishment from the central warehouse is carried out. It is assumed that the central warehouse has ample stock, so the model is a single-echelon model. Constraints are assumed on the aggregate mean waiting time, i.e., the average waiting time for

an arbitrary request. Under these constraints, they minimize expected cost, consisting of holding and transportation cost for lateral and emergency shipments. To obtain a lower bound on the optimal cost for the multi-item problem, they use Lagrangian relaxation to separate the problem into multiple single-item optimization problems, and they use a sub-gradient method to find the optimal values for the Lagrange multipliers (there are two multipliers since there are two service level constraints, one for each local warehouse). Starting from this lower bound, a feasible solution, i.e. an upper bound on the optimal cost, is obtained by means of a greedy procedure followed by local search. Besides a comparison to the cost performance under a single-item approach, they provide results on the potential savings of the application of lateral transshipment. Further, they show that the gap between the lower and upper bound is tight, indicating that the quality of their solution is good. In Wong et al. (2005), they extend their model to more than two local warehouses. Again Lagrangian relaxation is used to obtain a lower bound, and they test the performance of four different heuristics against this lower bound.

There is a close relationship between the model of Wong et al. (2005) and our model. Both are multi-item models for the multi-location situation with lateral transshipment, and both apply waiting time constraints. The main differences are that Wong et al. assume full pooling while we have a network structure that allows for partial pooling as well, and that Wong et al. apply exact evaluation while we use approximate evaluation (because of the number of local warehouses in real-life instances). Furthermore, Wong et al. have one waiting time constraint per local warehouse, while we allow multiple groups that each have a constraint on the aggregate mean waiting time. In fact, this means that we allow for commonality in our model as well. Lastly, Wong et al. use a different rule to determine which local warehouse supplies the lateral transshipment.

The approximate evaluation method that we provide is related to the approximate evaluation methods in Axsäter (1990), Alfredsson and Verrijdt (1999), Kukreja (2001), and Kutanoglu (2008). Axsäter studies a two-echelon system with lateral transshipment (between local warehouses in the same pooling group; see below where we discuss different forms of partial pooling), but without emergency shipment. I.e., a demand at a local warehouse that cannot be satisfied from stock or via a lateral transshipment, is simply backordered at the local warehouse. Axsäter develops an iterative, approximate evaluation algorithm based on a decomposition of the connected local warehouses into individual local warehouses and a modeling of the overflow demand streams (due to lateral transshipment) as Poisson demand streams with appropriate rates. The individual local warehouses are modeled as $M|G|\infty$ queues with state-dependent arrival rates. A similar approach was followed by Kukreja et al. for a single-echelon variant of Axsäter’s model. Alfredsson and Verrijdt studied a two-echelon system with both lateral transshipment and emergency shipment. They follow a similar approach as Axsäter for the local warehouses, where the local warehouses are modeled as Erlang loss systems ($M|G|c|c$ queues) with state-independent arrival rates. For a single-echelon variant of

the model of Alfredsson and Verrijdt, Kutanoglu extended their approximate evaluation method to determine time-based fill rates. In our single-echelon model, we have both lateral transshipment and emergency shipment as well. Similar to Alfredsson and Verrijdt, our approximate evaluation method is based on a modeling of the individual local warehouses as Erlang loss systems with state-independent arrival rates. However, we model the overflow demand streams differently. In our case, a lateral transshipment is asked from other local warehouses according to a deterministic instead of random order and we do *not* first check which other local warehouses have stock available. As a result, our iterative algorithm is based on different equations.

Finally, in our model, we allow *partial pooling*. Some forms of partial pooling have been studied in spare parts literature, but they differ from our form of partial pooling. A first form has been studied in Lee (1987) and Axsäter (1990). Both assume pooling groups, established on the basis of geographical proximity. Within those groups full pooling is applied, but no pooling between groups is allowed. Thus, this can be seen as partial pooling in the sense that inventory is shared with only *a few of the other locations*. Another form of partial pooling is that only *part of the inventory* is shared with other locations. Tagaras and Cohen (1992) study a two-location problem where a location can hold back stock and thus only shares a part of its inventory (in this paper, also pro-active lateral transshipment is allowed). A second reference regarding this form of partial pooling is Axsäter (2003b), who provides a heuristic decision rule to determine if and how many parts should be provided by means of lateral transshipment. Other papers with this form of partial pooling are Evers (2001), Minner et al. (2003), and Xu et al. (2003). Also, in networks with decentralized control, individual players are sometimes only willing to share part of their inventory with others (see e.g. Zhao et al., 2005). A third form of partial pooling is explored in Archibald et al. (1997) and Archibald (2007): in a periodic review setting, reactive lateral transshipment between locations can be applied. Archibald et al. (1997) explore a two-location problem, and Archibald (2007) explores a multi-location problem. The decision to use either lateral transshipment or emergency transshipment is dependent on the current stock level and the time remaining until the next replenishment. Pooling is thus applied partially, which is interpreted here as *in a fraction of the out-of-stock occurrences, depending on the state of the system*.

Contrary to all the above forms of partial pooling, we use the term *partial pooling* to indicate that only *part of the locations have the ability to act as provider of a lateral transshipment*. To the best of our knowledge, partial pooling in the way we define it has not been studied earlier. One special case, namely with one main local warehouse and one regular local warehouse, could be interpreted as a situation with unidirectional lateral transshipment (see Axsäter, 2003a, for approximate evaluation of unidirectional lateral transshipment), but apart from that special case, partial pooling and unidirectional lateral transshipment are different.

The structure of this paper is as follows. In Section 2, we describe our model. In Section 3, we

Figure 1 Graphical representation of pooling structure with main and regular local warehouses

describe how a given policy, i.e., a given vector of base stock levels in all local warehouses, can be evaluated exactly. After that, we describe and test our approximate evaluation method in Section 4. Next, we describe our greedy (heuristic) method to find a feasible policy in Section 5, followed by a comparison between the performance of partial and full pooling in Section 6. Then, we present a case study at ASML in Section 7, and we conclude in Section 8.

2 Model description

Let N denote a (non-empty) set of groups of installed machines at customers. Machines in each group $n \in N$ consist of multiple items (or SKU-s). The SKU-s can break down, which causes the corresponding machine to be non-operational. When the defective part is replaced by a new one, the machine is operational again. Let I denote the (non-empty) set of SKU-s, numbered $i = 1, \dots, |I|$. Failure rates (demand rates) per time unit for each SKU $i \in I$ and group $n \in N$ are assumed to follow a Poisson process with a constant rate $m_{i,n}$. We define M_n as the total demand for group n , i.e., $M_n := \sum_{i \in I} m_{i,n}$, $n \in N$, and we assume that $M_n > 0$, $n \in N$. The target aggregate mean waiting time for group $n \in N$, i.e., the maximum expected waiting time for an arbitrary request from group n , is denoted by $\hat{W}_n^{\text{obj}}$. The OEM has to meet these specified constraints, and aims to do so against minimal cost.

Let J denote the (non-empty) set of local warehouses, numbered $j = 1, \dots, |J|$. We distinguish two types of local warehouses: *main local warehouses* and *regular local warehouses*, shortly referred to as mains and regulars, respectively. Let $K (\subseteq J)$ denote the subset of main local warehouses. All other local warehouses $j \in J \setminus K$ are regular local warehouses. The difference between these two types of local warehouses is that mains can and regulars cannot be the supplier of a lateral transshipment. Although in principle K can be empty, in our model description and analysis below we assume $|K| > 0$. If $|K| = 0$, no lateral transshipment takes place, and then the analysis is straightforward because it can be done for each regular local warehouse separately. A graphical representation of the the network is given in Figure 1.

Each group $n \in N$ is assigned to exactly one local warehouse $j \in J$ (either a main or a regular

local warehouse). This means that if the group needs a spare part because a part of a machine in this group has broken down, it will submit its request to that particular local warehouse. Let $N_j (\subseteq N)$ denote the subset of groups that is assigned to local warehouse $j \in J$; so, the sets $N_j, j \in J$, constitute a partition of N . A local warehouse can serve more than one group; a main local warehouse could also have zero groups. Let $M_{i,j}, i \in I, j \in J$, denote the total demand for SKU i at local warehouse j , i.e., $M_{i,j} := \sum_{n \in N_j} m_{i,n}$. If a part is requested by group $n \in N_j$, it will be provided immediately by local warehouse j if this local warehouse has stock on hand. The corresponding waiting time is zero, which also holds for the corresponding cost.

If local warehouse j (either a regular or a main local warehouse) does not have stock on hand, then it tries to obtain the part by means of a lateral transshipment from one of the (other) main local warehouses. The corresponding transportation time for this lateral transshipment from main $k \in K, k \neq j$, to local warehouse $j \in J$ is $t_{j,k}^{\text{lat}} (\geq 0)$ and the corresponding cost is $C_{j,k}^{\text{lat}} (\geq 0)$. Notice that j can be both a main or a regular local warehouse. In case of a need for lateral transshipment, the main local warehouses are checked in a pre-specified order, and the first main that has stock on hand delivers the part. The pre-specified order follows the following structure. Each regular local warehouse is assigned to one main, i.e., in case of a need for lateral transshipment, this main is checked first. A main can have multiple regulars assigned to it. The main local warehouse $k \in K$ to which a regular local warehouse $j \in J \setminus K$ is assigned, is denoted as k_j . Furthermore, for each main local warehouse $k \in K$, a sequence is given for all other mains to be checked. Thus, the sequence of mains to be checked is identical for demands observed at main $k \in K$ and demands observed at a regular $j \in J \setminus K$ with $k_j = k$. The pre-specified order can be such that you have increasing transportation times, but other orderings are possible as well. Define vector $\sigma(k) := (\sigma_1(k), \dots, \sigma_{|K|-1}(k))$ as the permutation of main local warehouses $K \setminus \{k\}$ that represents this pre-specified order for main local warehouse k (each main local warehouse other than k appears exactly once in this order). Furthermore, let $K(k, \tilde{k}) (\subset K)$ denote the subset of main local warehouses with a lower position number than main $\tilde{k}$ in the pre-specified order for main local warehouse k , i.e., the subset containing all predecessors of $\tilde{k}$ in the pre-specified order for main local warehouse k . Notice that $K(k, \sigma_1(k)) = \emptyset$.

If neither (main or regular) local warehouse j itself nor one of the (other) main local warehouses can deliver the part, an emergency replenishment from a central warehouse takes place. The corresponding transportation time and cost (additional cost compared to normal replenishment) are $t_j^{\text{em}} (\geq t_{j,k}^{\text{lat}}, k \in K, k \neq j)$, and $C_j^{\text{em}} (\geq C_{j,k}^{\text{lat}}, k \in K, k \neq j)$, respectively. We assume that the central warehouse has infinite stock of all SKU-s and thus can always deliver requested parts.

The stock in all local warehouses is controlled by a base stock policy. The base stock level for SKU $i \in I$ in local warehouse $j \in J$ is denoted as $S_{i,j} (\in \mathbb{N}_0 = \mathbb{N} \cup \{0\})$. Let $S_i := (S_{i,1}, \dots, S_{i,|J|})$, $i \in I$, denote the vector of base stock levels for SKU i . Once a part in a local warehouse is used

to satisfy demand, immediately a new part is requested from the central warehouse. This part will be delivered after a certain normal replenishment lead time with mean $t_j^{\text{reg}} (\geq t_j^{\text{em}}, j \in J)$. No cost is involved here for transportation. (We could easily incorporate such a cost, but we omit it here since it represents a constant factor.) The replenishment lead times are assumed to be exponentially distributed and the means are the same for all local warehouses. The assumption of equal means is explicitly used in the approximate evaluation method, while this assumption is easily relaxed in the exact evaluation method. It is a justified assumption because in many spare parts networks in practice replenishment lead times are simply fixed at the same values for all local warehouses in the same geographical area. The assumption of an exponential distribution is made for the following reason. Alfredsson and Verrijdt (1999) studied a two-echelon model with lateral transshipment, and they have shown by simulation that their model is to a large extent insensitive to the replenishment lead time distribution for the local warehouses. It is reasonable to assume that this also holds for our model. The assumption of exponential replenishment lead times is explicitly used by the exact evaluation method, while the approximate evaluation method is also applicable for generally distributed replenishment lead times.

The cost for holding one unit of SKU $i \in I$ in stock for one time unit is C_i^h . We do assume that holding cost is also incurred for parts in replenishment. This is a logical choice since in general the whole spare parts supply chain is the property of one company.

The network under study is a single-echelon, multi-location network, since we assume infinite stock at the central warehouse. For ease of notation, we assume that transportation cost and transportation time parameters do not differ per SKU $i \in I$. However, it is easy to extend the model in this regard. We assume that t_j^{em} is considerably larger than $t_{j,k}^{\text{lat}}, j \in J, k \in K, k \neq j$, and that C_j^{em} is considerably larger than $C_{j,k}^{\text{lat}}, j \in J, k \in K, k \neq j$. Then, lateral transshipment is preferred over emergency replenishment (see also Kranenburg, 2006, Chapter 5).

With respect to the fulfillment of demand for SKU $i \in I$ at local warehouse $j \in J$, we introduce the following notation:

- $\beta_{i,j}(S_i)$, for the fraction of the demand for SKU i at local warehouse j that is delivered immediately upon request, i.e., from the stock in local warehouse j itself, also called the *(item) fill rate*;
- $\alpha_{i,j,k}(S_i), k \in K, k \neq j$, for the fraction of the demand for SKU i at local warehouse j that is delivered from main local warehouse k by means of lateral transshipment;
- $\theta_{i,j}(S_i)$, for the fraction of the demand for SKU i at local warehouse j that is delivered from the central warehouse (infinite stock) as an emergency shipment.

Besides the notation introduced above, define

$$A_{i,j}(S_i) := \sum_{k \in K, k \neq j} \alpha_{i,j,k}(S_i)$$

as the total fraction of the demand for SKU $i \in I$ at local warehouse $j \in J$ that is delivered by lateral transshipment. Notice that for each SKU $i \in I$ at each local warehouse $j \in J$, it holds that

$$\beta_{i,j}(S_i) + A_{i,j}(S_i) + \theta_{i,j}(S_i) = 1. \quad (1)$$

Furthermore, notice that $\theta_{i,k}(S_i)$ is equal for all mains $k \in K$ because an emergency transshipment from the central warehouse only takes place if none of the main local warehouses has stock on hand, and, obviously, this probability is independent on which main local warehouse observes the demand.

In Sections 3 and 4, we show how all fractions can be determined, in an exact and approximate way, respectively, for a given vector of base stock levels S_i , $i \in I$.

Let $W_{i,j}(S_i)$, $i \in I$, $j \in J$, denote the expected waiting time if SKU i is requested at local warehouse j , at given vector S_i . Then, $W_{i,j}(S_i)$ can be calculated as follows:

$$W_{i,j}(S_i) = \sum_{k \in K, k \neq j} t_{j,k}^{\text{lat}} \alpha_{i,j,k}(S_i) + t_j^{\text{em}} \theta_{i,j}(S_i). \quad (2)$$

For SKU $i \in I$, the total holding cost incurred per time unit equals

$$\sum_{j \in J} C_i^h S_{i,j}, \quad (3)$$

and the expected total transportation cost incurred per time unit equals

$$\sum_{j \in J} M_{i,j} \left(\sum_{k \in K, k \neq j} C_{j,k}^{\text{lat}} \alpha_{i,j,k}(S_i) + C_j^{\text{em}} \theta_{i,j}(S_i) \right).$$

Therefore, for SKU $i \in I$, the expected total cost per time unit, $C_i(S_i)$, is

$$C_i(S_i) = \sum_{j \in J} C_i^h S_{i,j} + \sum_{j \in J} M_{i,j} \left(\sum_{k \in K, k \neq j} C_{j,k}^{\text{lat}} \alpha_{i,j,k}(S_i) + C_j^{\text{em}} \theta_{i,j}(S_i) \right). \quad (4)$$

The objective of the OEM is to minimize the expected total cost for all SKU-s together, under the condition that the expected waiting time for an arbitrary request from each group $n \in N$ does not exceed the target aggregate mean waiting time $\hat{W}_n^{\text{obj}}$. Since the expected waiting time for an arbitrary request from group $n \in N_j$ is a demand-weighted average of the expected waiting times of all SKU-s at local warehouse j , the problem can be formulated mathematically as follows:

$$\begin{aligned} \text{(P)} \quad & \min \quad \sum_{i \in I} C_i(S_i) \\ & \text{subject to} \quad \sum_{i \in I} \frac{m_{i,n}}{M_n} W_{i,j}(S_i) \leq \hat{W}_n^{\text{obj}}, \quad j \in J, n \in N_j, \\ & S_{i,j} \in \mathbb{N}_0, \quad i \in I, j \in J. \end{aligned}$$

In the next sections, Sections 3, 4, and 5, we describe how a given policy can be evaluated, either exactly or approximately, and how a feasible solution for Problem (P) can be obtained.

3 Exact evaluation

In this section, we describe how a certain policy, i.e., a choice of base stock levels for all main and regular local warehouses, can be evaluated exactly. Evaluation can be done for each SKU $i \in I$ separately. In the evaluation, for a given policy S_i for SKU i , i.e., for a given vector of base stock levels in all local warehouses $j \in J$ for SKU i , we determine $\beta_{i,j}(S_i)$, $j \in J$, $\alpha_{i,j,k}(S_i)$, $j \in J$, $k \in K$, $k \neq j$, and $\theta_{i,j}(S_i)$, $j \in J$. Once we have determined those, the waiting times $W_{i,j}(S_i)$, $j \in J$, and cost $C_i(S_i)$, follow immediately from Equations (2) and (4).

For the exact evaluation of a given policy for SKU $i \in I$, we may use a Markov process description. The states may be described by the on-hand stocks at all the warehouses $j \in J$. This gives a $|J|$ -dimensional state space, and thus exact numerical evaluation can only be done for a limited number of local warehouses, as computation times are exponential in the number of local warehouses.

For our Markov process with a $|J|$ -dimensional state space, let $x_i = (x_{i,1}, \dots, x_{i,|J|})$ denote a vector that represents the current state, where $x_{i,j}$, $j \in J$, is the number of parts on hand in local warehouse j , and with $0 \leq x_{i,j} \leq S_{i,j}$, $j \in J$. Furthermore, let e_j denote a row vector of size $|J|$ with the j -th element equal to 1 and all other elements equal to 0.

If the system is in state x_i , then two types of events can occur. First, at any local warehouse $j \in J$, a request can arrive for SKU i . Secondly, at each local warehouse $j \in J$ that has less stock on hand than its base stock level ($x_{i,j} < S_{i,j}$), a part of SKU i can arrive as a stock replenishment from the central warehouse. By considering both types of events for all applicable local warehouses, we obtain the transition rates from state x_i to other states. To identify all transitions, this has to be done for all states.

If a part of SKU i is requested at a regular local warehouse $j \in J \setminus K$ while the system is in state x_i , then one of the following cases is true:

- Regular local warehouse j has stock on hand. In this case, the regular local warehouse can supply the requested part from its own stock. The corresponding transition is $x_i \rightarrow x_i - e_j$, and the transition rate is $M_{i,j}$.
- Regular local warehouse j has no stock on hand, but main local warehouse k_j has stock on hand. In this case, main local warehouse k_j delivers the requested part. The corresponding transition is $x_i \rightarrow x_i - e_{k_j}$, and the transition rate is $M_{i,j}$.
- Neither regular local warehouse j nor main local warehouse k_j has stock on hand, but at least one other main local warehouse has stock on hand. In this case, the part is delivered by

the first main local warehouse k in the predefined lateral transshipment order $\sigma(k_j)$ that has stock on hand. The corresponding transition is $x_i \rightarrow x_i - e_k$, and the transition rate is $M_{i,j}$.

- Neither regular local warehouse j nor any of the main local warehouses has stock on hand. In this case, the part will be supplied by an emergency replenishment from the central warehouse, and the system remains in the current state (no transition takes place).

If a part of SKU i is requested at a main local warehouse $j \in K$ while the system is in state x_i , then one of the following cases is true:

- Main local warehouse j has stock on hand. In this case, the main local warehouse can supply the requested part from its own stock. The corresponding transition is $x_i \rightarrow x_i - e_j$, and the transition rate is $M_{i,j}$.
- Main local warehouse j has no stock on hand, but at least one other main local warehouse has stock on hand. In this case, the part is delivered by the first main local warehouse k in the predefined lateral transshipment order $\sigma(j)$ that has stock on hand. The corresponding transition is $x_i \rightarrow x_i - e_k$, and the transition rate is $M_{i,j}$.
- None of the main local warehouses has stock on hand. In this case, the part will be supplied by an emergency replenishment from the central warehouse, and the system remains in the current state (no transition takes place).

If a part of SKU i arrives as stock replenishment in local warehouse $j \in J$ with $x_{i,j} < S_{i,j}$ while the system is in state x_i , then the transition is $x_i \rightarrow x_i + e_j$, and the transition rate is $(S_{i,j} - x_{i,j})/t^{\text{reg}}$. The numerator is the number of parts of local warehouse j that is currently being replenished.

When all transitions are determined, we can apply uniformization (see e.g. Ross, 1996, Section 5.8) by adding transitions from a state to itself. After uniformization, we can find the equilibrium distribution, i.e., the steady-state probabilities, for our Markov process by successive approximation, also known as the power method (see e.g. Varga, 1962, p. 281). Theoretically, successive approximation provides us with exact steady-state probabilities. Numerically, however, we will stop when steady-state probabilities do not change more than ϵ , with ϵ small. For all states x_i , let $\pi(x_i)$ denote the steady-state probability that the Markov process is in state x_i .

The next step is to determine $\beta_{i,j}(S_i)$, $\alpha_{i,j,k}(S_i)$, $k \in K$, $k \neq j$, and $\theta_{i,j}(S_i)$, for all local warehouses $j \in J$. Remember that for each local warehouse j , these values sum up to 1. Each steady-state probability $\pi(x_i)$ contributes to exactly one of the terms $\beta_{i,j}(S_i)$, $\alpha_{i,j,k}(S_i)$, and $\theta_{i,j}(S_i)$, and thus they can be determined by summation of the correct probabilities. For each regular local

warehouse $j \in J \setminus K$, it holds that

$$\begin{aligned}
\beta_{i,j}(S_i) &= \sum_{x_i | x_{i,j} > 0} \pi(x_i); \\
\alpha_{i,j,k_j}(S_i) &= \sum_{x_i | x_{i,j} = 0, x_{i,k_j} > 0} \pi(x_i); \\
\alpha_{i,j,k}(S_i) &= \sum_{x_i | x_{i,j} = 0, x_{i,k_j} = 0, x_{i,\hat{k}} = 0 \ (\forall \hat{k} \in K(k_j, k)), x_{i,k} > 0} \pi(x_i), \quad k \in K, k \neq k_j; \\
\theta_{i,j}(S_i) &= \sum_{x_i | x_{i,j} = 0, x_{i,k} = 0, (\forall k \in K)} \pi(x_i).
\end{aligned}$$

For each main local warehouse $j \in K$, we obtain:

$$\begin{aligned}
\beta_{i,j}(S_i) &= \sum_{x_i | x_{i,j} > 0} \pi(x_i); \\
\alpha_{i,j,k}(S_i) &= \sum_{x_i | x_{i,j} = 0, x_{i,\hat{k}} = 0 \ (\forall \hat{k} \in K(j, k)), x_{i,k} > 0} \pi(x_i), \quad k \in K, \quad k \neq j; \\
\theta_{i,j}(S_i) &= \sum_{x_i | x_{i,k} = 0, (\forall k \in K)} \pi(x_i).
\end{aligned}$$

Of course, for each local warehouse $j \in J$, one of the terms $\beta_{i,j}(S_i)$, $\alpha_{i,j,k}(S_i)$, and $\theta_{i,j}(S_i)$, could be calculated alternatively as the difference between 1 and the sum of the other terms.

4 Approximate evaluation

In Section 3, we have described how a base stock policy can be evaluated exactly. However, exact evaluation is done numerically, and can be time-consuming or even computationally intractable if the number of local warehouses is large, since each local warehouse constitutes a dimension in the Markov process. (Of course, in the special case that all local warehouses are regulars, i.e., if no lateral transshipment takes place at all, each regular can be analyzed individually and we do not have computational problems.) Since we want our method to be applicable for real-life instances with many local warehouses, we introduce an approximate evaluation method in this section.

As mentioned in the previous section, evaluation can be done for each SKU $i \in I$ separately. In the approximate evaluation, for a given policy S_i for SKU i , we approximate the $\beta_{i,j}(S_i)$, $\alpha_{i,j,k}(S_i)$, and $\theta_{i,j}(S_i)$; the waiting times $W_{i,j}(S_i)$ and cost $C_i(S_i)$ then follow from Equations (2) and (4).

In the approximate evaluation, we make use of the loss probability in the Erlang loss model (i.e., the $M|G|c|c$ queue). Let $L(c, \rho)$ denote this loss probability, where c represents the number of servers and ρ the offered load. Then, $L(c, \rho)$ is given by (see e.g. Tijms, 2003)

$$L(c, \rho) = \frac{\rho^c / c!}{\sum_{x=0}^c \rho^x / x!}.$$

The key to our approximate evaluation method is that: (i) We decouple the network into individual local warehouses; (ii) We assume that the extra demand processes in main local warehouses due to requests for lateral transshipment are Poisson processes. The decoupling is done in two

steps. In the first step, we decouple the regular local warehouses from the mains. In the second step the resulting system of main local warehouses is decoupled, so that each main can be analyzed individually. These two decoupling steps are discussed in §4.1 and §4.2, respectively. After that, we describe our approximate evaluation method more formally in §4.3. In §4.4, we compare our approximate evaluation to exact evaluation.

4.1 Decoupling the regulars from the mains

The connection between regular local warehouses $j \in J \setminus K$ and all main local warehouses $k \in K$ is that the mains can provide a lateral transshipment to a regular. A demand for such a lateral transshipment occurs when a regular j faces customer demand at a moment that it is out of stock. We refer to the demand process for lateral transshipment as generated by regular $j \in J \setminus K$ as an *overflow demand process*. In our approximate evaluation, *we assume that the overflow demand process at regular $j \in J \setminus K$ behaves as a Poisson process that constitutes an additional demand stream at main k_j* . Obviously, in reality this demand process can be burstier than Poisson, as it occurs only at the moments that the regular local warehouse is out of stock, but for low demand rates our assumption seems reasonable. Also, notice that in case a regular j has $S_{i,j} = 0$, the overflow demand really follows a Poisson process, since all demand is forwarded.

First, each regular $j \in J \setminus K$ is analyzed separately using the Erlang loss model. The steady-state behavior of the stock in a regular local warehouse $j \in J \setminus K$ is identical to the steady-state behavior of the number of idle servers in an Erlang loss system with demand rate $M_{i,j}$, $S_{i,j}$ servers, and the mean replenishment time t^{reg} as mean service time. Hence, $\beta_{i,j}(S_i)$ (which actually only depends on $S_{i,j}$, and not on the base stock levels at other local warehouses) is equal to the corresponding Erlang loss probability, i.e., $\beta_{i,j}(S_i) = 1 - L(S_{i,j}, M_{i,j}t^{\text{reg}})$ (so, this formula is exact).

At this point, we use our assumption that the demand for lateral transshipment placed by regular j at main k_j follows a Poisson process; the corresponding rate is $(1 - \beta_{i,j}(S_i))M_{i,j}$. We introduce $\widetilde{M}_{i,k}$ for the demand rate for SKU i at main local warehouse $k \in K$ *including* all demand for lateral transshipment from regulars assigned to this main, i.e., $\widetilde{M}_{i,k} := M_{i,k} + \sum_{j \in J | k_j = k} (1 - \beta_{i,j}(S_i)) M_{i,j}$. We can now analyze the system of mains, where each main $k \in K$ faces a Poisson demand process with parameter $\widetilde{M}_{i,k}$; this analysis is described in §4.2. The analysis of the system of mains provides us with $\beta_{i,k}(S_i)$, $k \in K$, $\alpha_{i,k,\tilde{k}}(S_i)$, $k \in K$, $\tilde{k} \in K$, $\tilde{k} \neq k$, and $\theta_{i,k}(S_i)$, $k \in K$.

Finally, we can determine the remaining performance measures $\alpha_{i,j,k}(S_i)$, $k \in K$, and $\theta_{i,j}(S_i)$ for all regulars $j \in J \setminus K$, based on the performance measures that we determined for the mains:

$$\alpha_{i,j,k}(S_i) := \begin{cases} (1 - \beta_{i,j}(S_i))\beta_{i,k_j}(S_i), & k = k_j, \\ (1 - \beta_{i,j}(S_i))\alpha_{i,k_j,k}(S_i), & k \in K, k \neq k_j, \end{cases} \quad (5)$$

and $\theta_{i,j}(S_i) := (1 - \beta_{i,j}(S_i))\theta_{i,k_j}(S_i)$.

Notice that, under the above assumption for the overflow demand processes, the system of mains could be analyzed exactly, by the method described in Section 3, in which case the computation time is determined by the numerical solution of the steady-state equations of a $|K|$ -dimensional Markov process. This may be attractive for low values of $|K|$ (≤ 3 , say). For larger $|K|$, a further decomposition is desired, which we describe in the next subsection.

4.2 Decoupling the mains

In the second decoupling step, we decouple the main local warehouses, so that each main can be analyzed separately. The connection between the main local warehouses $k \in K$ is that lateral transshipment can take place from each main to each other main. A demand for a lateral transshipment occurs when a main $k \in K$ faces customer demand at a moment that it is out of stock. We refer to the demand process for lateral transshipment arising at main $k \in K$ as the *overflow demand process* at main $k \in K$. In our approximate evaluation, *we assume that the overflow demand process at main $k \in K$ behaves as a Poisson process that constitutes additional demand streams at all other mains l , $l \in K$, $l \neq k$* . Again, in reality this demand process could be burstier than Poisson, as it occurs only at the moments that main k is out of stock. However, by making the assumption that it follows a Poisson process, we can decouple the mains, and analyze each main individually. For the analysis of each main local warehouse, we use the Erlang loss model.

Recall that the demand processes at the main local warehouses $k \in K$ are approximated by Poisson processes with rates $\widetilde{M}_{i,k}$ (where the overflow processes from the regular local warehouses are included) and they have base stock levels $S_{i,k}$. First, under the assumption that these demand processes are exactly equal to Poisson processes (we have this in special cases, e.g. when we have a system with mains only), we can determine the $\theta_{i,k}(S_i)$ exactly by considering the *aggregate* system of all main local warehouses $k \in K$ together. Then, because of full pooling of inventory between the main local warehouses and the application of emergency shipments in case all main local warehouses are out of stock, the steady-state behavior of the total stock in the aggregate system is identical to the steady-state behavior of the number of idle servers in an Erlang loss system with demand rate $\sum_{k \in K} \widetilde{M}_{i,k}$, $\sum_{k \in K} S_{i,k}$ servers, and the mean replenishment time t^{reg} as mean service time. Thus, we can calculate $\theta_{i,k}(S_i)$, $k \in K$, as the Erlang loss probability of the aggregate system:

$$\theta_{i,k}(S_i) := L \left(\sum_{k \in K} S_{i,k}, \sum_{k \in K} \widetilde{M}_{i,k} t^{\text{reg}} \right), \quad k \in K.$$

So, in special cases, this expression is exact. In general, the demand processes at the main local warehouses $k \in K$ will be almost equal to Poisson processes, but not exactly, and then this expression is an approximation.

At this point, we introduce $\widehat{M}_{i,\widetilde{k},k}$, $k, \widetilde{k} \in K$, $\widetilde{k} \neq k$, for the rate with which main $\widetilde{k}$ requests a lateral transshipment from main k . Furthermore, we introduce $\widehat{M}_{i,k}$ for the demand rate for SKU

i at main local warehouse $k \in K$ including *all* demand for lateral transshipment. Thus,

$$\widehat{M}_{i,k} := \widetilde{M}_{i,k} + \sum_{\widetilde{k} \in K, \widetilde{k} \neq k} \widehat{M}_{i,\widetilde{k},k}, \quad k \in K.$$

In our approximate evaluation, we analyze each main local warehouse separately, using the Erlang loss model. For each main $k \in K$, $\beta_{i,k}(S_i) := 1 - L\left(S_{i,k}, \widehat{M}_{i,k} t^{\text{reg}}\right)$, and $A_{i,k}(S_i) := 1 - (\beta_{i,k}(S_i) + \theta_{i,k}(S_i))$. So, $\beta_{i,k}(S_i)$ and $A_{i,k}(S_i)$ are dependent on $\widehat{M}_{i,k}$. (It is discussed later how we obtain the values for $\alpha_{i,k,\widetilde{k}}(S_i)$, $\widetilde{k} \in K$, $\widetilde{k} \neq k$.)

On the other hand, $\widehat{M}_{i,k}$, $k \in K$, is dependent on fill rates. We determine its components $\widehat{M}_{i,\widetilde{k},k}$, $\widetilde{k} \in K$, $\widetilde{k} \neq k$, as:

$$\widehat{M}_{i,\widetilde{k},k} := \begin{cases} \frac{A_{i,\widetilde{k}}(S_i) \widetilde{M}_{i,\widetilde{k}}}{1 - \prod_{\ell \in K, \ell \neq \widetilde{k}} (1 - \beta_{i,\ell}(S_i))} \prod_{\ell \in K(\widetilde{k},k)} (1 - \beta_{i,\ell}(S_i)), & S_{i,\ell} > 0 \text{ for at least} \\ & \text{one } \ell \in K \setminus \{\widetilde{k}\}, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

with the product term $\prod_{\ell \in K(\widetilde{k},k)} (1 - \beta_{i,\ell}(S_i))$ defined as 1 if $K(\widetilde{k},k) = \emptyset$. The explanation for this formula follows below.

We explain Equation (6) by means of an example. Consider a system with 3 mains, numbered 1, 2, and 3, with $\sigma(1) = (2, 3)$. We derive Equation (6) for $\widehat{M}_{i,1,2}$ and $\widehat{M}_{i,1,3}$ for the case that $S_{i,2} > 0$ or $S_{i,3} > 0$. In the Erlang loss model for main 2, a fraction $\beta_{i,2}(S_i)$ of the observed Poisson demand rate $\widehat{M}_{i,1,2}$ for lateral transshipment as requested by main 1 will be fulfilled. According to the pre-defined lateral transshipment order $\sigma(1)$, we let main 3 face the remaining demand for lateral transshipment to main 1, i.e., $\widehat{M}_{i,1,3} := (1 - \beta_{i,2}(S_i)) \widehat{M}_{i,1,2}$, again assuming that this demand follows a Poisson process. In the Erlang loss model for main 3, a fraction $\beta_{i,3}(S_i)$ of the observed Poisson demand rate $\widehat{M}_{i,1,3}$ for lateral transshipment as requested by main 1 will be fulfilled. Under the assumption of independency of the physical stocks in mains 2 and 3, we have that (in expectation) in total the following amount is satisfied by means of lateral transshipment:

$$\begin{aligned} & \beta_{i,2}(S_i) \widehat{M}_{i,1,2} + \beta_{i,3}(S_i) \widehat{M}_{i,1,3} \\ &= \beta_{i,2}(S_i) \widehat{M}_{i,1,2} + \beta_{i,3}(S_i) (1 - \beta_{i,2}(S_i)) \widehat{M}_{i,1,2} \\ &= (\beta_{i,2}(S_i) + \beta_{i,3}(S_i) - \beta_{i,2}(S_i) \beta_{i,3}(S_i)) \widehat{M}_{i,1,2} \\ &= [1 - (1 - \beta_{i,2}(S_i))(1 - \beta_{i,3}(S_i))] \widehat{M}_{i,1,2}. \end{aligned} \quad (7)$$

Because in total a fraction $A_{i,1}(S_i)$ of the demand $\widetilde{M}_{i,1}$ at main 1 is supplied by lateral transshipment from the other mains 2 and 3, Equation (7) equals $A_{i,1}(S_i) \widetilde{M}_{i,1}$, or, equivalently,

$$\widehat{M}_{i,1,2} = \frac{A_{i,1}(S_i) \widetilde{M}_{i,1}}{1 - (1 - \beta_{i,2}(S_i))(1 - \beta_{i,3}(S_i))}. \quad (8)$$

And thus,

$$\widehat{M}_{i,1,3} = \frac{A_{i,1}(S_i)\widetilde{M}_{i,1}}{1 - (1 - \beta_{i,2}(S_i))(1 - \beta_{i,3}(S_i))}(1 - \beta_{i,2}(S_i)). \quad (9)$$

Equations (8) and (9) are in accordance with Equation (6).

If $S_{i,\ell} = 0$ for all $\ell \in K \setminus \tilde{k}$, then $\widehat{M}_{i,\tilde{k},k} = 0$. We distinguish this case since then the denominator $1 - \prod_{\ell \in K, \ell \neq \tilde{k}} (1 - \beta_{i,\ell}(S_i))$ in Equation (6) would be zero. Notice that in this case, no lateral transshipment to main $\tilde{k}$ will take place at all.

We use an *iterative procedure* to determine $\widehat{M}_{i,k}$, $k \in K$, $\widehat{M}_{i,\tilde{k},k}$, $k, \tilde{k} \in K$, $\tilde{k} \neq k$, $\beta_{i,k}(S_i)$, $k \in K$, and $A_{i,k}(S_i)$, $k \in K$. This procedure works as follows. Initially, we suppose that no lateral transshipment takes place between the mains and accordingly, we set the demand rates equal to $\widehat{M}_{i,k} := \widetilde{M}_{i,k}$, $k \in K$. We apply the Erlang loss model for each main $k \in K$ and determine $\beta_{i,k}(S_i) := 1 - L(S_{i,k}, \widehat{M}_{i,k} t^{\text{reg}})$, and $A_{i,k}(S_i) := 1 - (\beta_{i,k}(S_i) + \theta_{i,k}(S_i))$, $k \in K$. For *one* main k' , we now determine $\widehat{M}_{i,\tilde{k},k'}$, $\tilde{k} \in K$, $\tilde{k} \neq k'$, using Equation (6), and, subsequently, $\widehat{M}_{i,k'}$, $\beta_{i,k'}(S_i)$, and $A_{i,k'}(S_i)$. We repeat this for all other mains $k \in K$, and then consider main k' again, etcetera. We continue this iterative procedure until the $\widehat{M}_{i,k}$, $k \in K$, each do not change more than ϵ , with ϵ small. We observed that this iterative procedure converges in all cases, but there is no formal proof for the convergence. After convergence, the values for $\alpha_{i,k,\tilde{k}}(S_i)$, $k \in K$, $\tilde{k} \in K$, $\tilde{k} \neq k$, are determined as $\alpha_{i,k,\tilde{k}}(S_i) := \beta_{i,\tilde{k}}(S_i) \widehat{M}_{i,k,\tilde{k}} / \widehat{M}_{i,k}$.

4.3 Formal description of the approximate evaluation method

In §4.1 and §4.2, we have described two decoupling steps that we use in our approximate evaluation method. In the current subsection, we describe the whole method formally. We describe two algorithms, Algorithms 1 and 2. Algorithm 2 describes Step 3 of Algorithm 1 in detail.

Algorithm 1

Step 1 For all regulars $j \in J \setminus K$, $\beta_{i,j}(S_i) := 1 - L(S_{i,j}, M_{i,j} t^{\text{reg}})$.

Step 2 For all mains $k \in K$, $\widetilde{M}_{i,k} := M_{i,k} + \sum_{j \in J | k_j = k} (1 - \beta_{i,j}(S_i)) M_{i,j}$.

Step 3 For all mains $k \in K$, determine $\beta_{i,k}(S_i)$, $\alpha_{i,k,\tilde{k}}(S_i)$, $\tilde{k} \in K$, $\tilde{k} \neq k$, and $\theta_{i,k}(S_i)$, using Algorithm 2.

Step 4 For all regulars $j \in J \setminus K$, if $K = \emptyset$, then $\theta_{i,j}(S_i) := (1 - \beta_{i,j}(S_i))$. Otherwise, $\alpha_{i,j,k}(S_i)$ is determined using Equation (5) and $\theta_{i,j}(S_i) := (1 - \beta_{i,j}(S_i))\theta_{i,k_j}(S_i)$.

Algorithm 2

Step 1 For all mains $k \in K$, $\theta_{i,k}(S_i) := L\left(\sum_{k \in K} S_{i,k}, \sum_{k \in K} \widetilde{M}_{i,k} t^{\text{reg}}\right)$.

Step 2 For all mains $k \in K$, $\beta_{i,k}(S_i) := 1 - L\left(S_{i,k}, \widetilde{M}_{i,k} t^{\text{reg}}\right)$, and
 $A_{i,k}(S_i) := 1 - (\beta_{i,k}(S_i) + \theta_{i,k}(S_i))$.

Step 3 For *one main* $k \in K$:

Step 3-a Determine $\widehat{M}_{i,\widetilde{k},k}$ using Equation (6), and

$$\widehat{M}_{i,k} := \widetilde{M}_{i,k} + \sum_{\widetilde{k} \in K, \widetilde{k} \neq k} \widehat{M}_{i,\widetilde{k},k}.$$

Step 3-b $\beta_{i,k}(S_i) := 1 - L\left(S_{i,k}, \widehat{M}_{i,k} t^{\text{reg}}\right)$, and

$$A_{i,k}(S_i) := 1 - (\beta_{i,k}(S_i) + \theta_{i,k}(S_i)).$$

Step 4 Repeat Step 3 for all other mains $k \in K$.

Step 5 Repeat Steps 3 and 4 until $\widehat{M}_{i,k}$ does not change more than ϵ for each $k \in K$,
 with ϵ small.

Step 6 For all mains $k \in K$, $\alpha_{i,k,\widetilde{k}}(S_i) := \beta_{i,\widetilde{k}}(S_i) \widehat{M}_{i,k,\widetilde{k}} / \widehat{M}_{i,k}$, $\widetilde{k} \in K$, $\widetilde{k} \neq k$.

Notice that in the full pooling situation, i.e., if all local warehouses are mains, our algorithm basically boils down to Algorithm 2. In the no pooling situation, i.e., if all local warehouses are regulars, then each regular is analyzed individually (using only Steps 1 and 4 of Algorithm 1), and this analysis is exact.

4.4 Numerical comparison

In this subsection we test our approximate evaluation method. For instances with a small number of local warehouses ($|J| \leq 4$) we compare the approximate evaluation to exact evaluation. As we mentioned, evaluation is done for each SKU individually, so our instances each consider one item only. We assume that each local warehouse serves one group, so we can represent demand rates as $M_{i,j}$, $j \in J$.

We first consider instances with main local warehouses only. Table 1 gives parameter settings and results for symmetric instances, and Tables 2 and 3 give parameter settings and results for asymmetric instances. Symmetric instances have equal demand rates and base stock levels in all mains, and have a cyclic pre-defined order for lateral transshipment, so that each main observes the same demand process for lateral transshipment. Asymmetric instances have different demand rates, different base stock levels per location, or both, and for these asymmetric instances, we also vary the pre-defined lateral transshipment order. We study instances with 2 and 4 main local warehouses, and different values for demand rates $M_{i,k}$ and base stock levels $S_{i,k}$, $k \in K$. Maximum values for demand rates match what we observe as maximum demand rates at ASML. Even for

Table 1 Parameter settings and results for symmetric instances with mains only, with $M_{i,k}$ in parts per year, with $t^{\text{reg}} = 0.04$ years, and with $\sigma(1) = (2)$ and $\sigma(2) = (1)$ for $|K| = 2$, and $\sigma(1) = (2, 3, 4)$, $\sigma(2) = (3, 4, 1)$, $\sigma(3) = (4, 1, 2)$ and $\sigma(4) = (1, 2, 3)$ for $|K| = 4$

$ K $	$M_{i,k}$	$S_{i,k}$	$\beta_{i,k}(S_i)$		$\alpha_{i,k,\sigma_1(k)}(S_i)$		$\alpha_{i,k,\sigma_2(k)}(S_i)$		$\alpha_{i,k,\sigma_3(k)}(S_i)$		$\theta_{i,k}(S_i)$
			exact	appr.	exact	appr.	exact	appr.	exact	appr.	
2	0.5	1	0.980	0.980	0.019	0.019					0.001
2	1	1	0.960	0.960	0.037	0.037					0.003
2	5	1	0.811	0.811	0.135	0.135					0.054
2	10	1	0.660	0.660	0.189	0.189					0.151
2	50	1	0.231	0.231	0.154	0.154					0.615
2	5	2	0.983	0.983	0.016	0.016					0.001
2	10	2	0.941	0.941	0.052	0.051					0.008
2	50	2	0.489	0.492	0.201	0.197					0.311
4	0.5	1	0.980	0.980	0.019	0.020	0.001	0.000	0.000	0.000	0.000
4	1	1	0.960	0.960	0.038	0.038	0.002	0.002	0.000	0.000	0.000
4	5	1	0.802	0.802	0.145	0.154	0.036	0.031	0.010	0.006	0.008
4	10	1	0.623	0.623	0.203	0.211	0.082	0.080	0.035	0.030	0.056
4	50	1	0.149	0.149	0.114	0.107	0.090	0.091	0.072	0.078	0.575
4	5	2	0.983	0.983	0.016	0.017	0.000	0.000	0.000	0.000	0.000
4	10	2	0.940	0.940	0.054	0.056	0.005	0.003	0.001	0.000	0.000
4	50	2	0.386	0.391	0.195	0.189	0.114	0.115	0.069	0.070	0.236

Table 2 Parameter settings for asymmetric instances with mains only, with $M_{i,k}$ in parts per year, with $t^{\text{reg}} = 0.04$ years, and with $\sigma(1) = (2)$ and $\sigma(2) = (1)$ for $|K| = 2$, and for $|K| = 4$ either $\sigma(1) = (2, 3, 4)$, $\sigma(2) = (3, 4, 1)$, $\sigma(3) = (4, 1, 2)$ and $\sigma(4) = (1, 2, 3)$ (cycle) or $\sigma(1) = (2, 3, 4)$, $\sigma(2) = (1, 3, 4)$, $\sigma(3) = (1, 2, 4)$ and $\sigma(4) = (1, 2, 3)$ (dom., i.e., dominance)

Instance	$ K $	$\sigma(k)$	$M_{i,1}$	$M_{i,2}$	$M_{i,3}$	$M_{i,4}$	$S_{i,1}$	$S_{i,2}$	$S_{i,3}$	$S_{i,4}$
1	2		1	5			1	1		
2	2		1	5			1	2		
3	2		5	10			1	1		
4	2		5	10			1	2		
5	4	cycle	1	5	5	10	1	1	1	1
6	4	cycle	1	5	5	10	1	1	1	2
7	4	cycle	1	5	5	10	1	1	2	2
8	4	cycle	1	5	5	10	1	2	2	2
9	4	cycle	5	5	5	5	1	1	1	2
10	4	cycle	5	5	5	5	1	2	2	2
11	4	dom.	1	5	5	10	1	1	1	1
12	4	dom.	1	5	5	10	1	1	1	2
13	4	dom.	1	5	5	10	1	1	2	2
14	4	dom.	1	5	5	10	1	2	2	2
15	4	dom.	5	5	5	5	1	1	1	2
16	4	dom.	5	5	5	5	1	2	2	2

Table 3 Results for asymmetric instances with mains only

Instance	$\beta_{i,1}(S_i)$		$\beta_{i,2}(S_i)$		$\beta_{i,3}(S_i)$		$\beta_{i,4}(S_i)$		$\theta_{i,k}(S_i)$
	exact	appr.	exact	appr.	exact	appr.	exact	appr.	
1	0.934	0.934	0.832	0.832					0.023
2	0.959	0.959	0.983	0.983					0.002
3	0.765	0.765	0.695	0.695					0.101
4	0.819	0.819	0.938	0.938					0.020
5	0.859	0.852	0.811	0.816	0.805	0.807	0.692	0.692	0.009
6	0.938	0.936	0.829	0.830	0.811	0.810	0.935	0.936	0.002
7	0.943	0.941	0.830	0.831	0.977	0.978	0.945	0.945	0.000
8	0.944	0.942	0.983	0.983	0.983	0.983	0.945	0.945	0.000
9	0.829	0.829	0.811	0.810	0.805	0.804	0.974	0.976	0.001
10	0.831	0.831	0.978	0.978	0.983	0.983	0.983	0.983	0.000
11	0.827	0.818	0.808	0.811	0.821	0.825	0.712	0.713	0.009
12	0.891	0.885	0.825	0.826	0.828	0.830	0.945	0.946	0.002
13	0.914	0.910	0.829	0.829	0.982	0.983	0.946	0.946	0.000
14	0.939	0.936	0.983	0.983	0.983	0.984	0.946	0.946	0.000
15	0.787	0.782	0.802	0.799	0.819	0.821	0.981	0.983	0.001
16	0.827	0.826	0.977	0.978	0.983	0.983	0.984	0.984	0.000

those maximum values, low base stock levels turn out to be sufficient to guarantee high fill rates. For this reason we limit ourselves to instances with base stock levels of at most 2.

For instances with main local warehouses only, $\theta_{i,k}(S_i)$, as determined in our approximate evaluation method, is equal for each $k \in K$. Furthermore, it is exact as it is obtained by an analysis of the aggregate system of all mains. Thus, the difference between the approximate and exact $\beta_{i,k}(S_i)$ is equal to the difference between the exact and approximate $A_{i,k}(S_i)$ (cf. (1)).

For symmetric instances, we report $\beta_{i,k}(S_i)$ as determined exactly and approximately, and $\theta_{i,k}(S_i)$. Further, we study the lateral transshipment fraction $A_{i,k}(S_i)$ on the disaggregated level by depicting $\alpha_{i,k,\tilde{k}}(S_i)$ (specifying which main $\tilde{k} \in K$ provides the lateral transshipment to main k). Table 1 shows that the maximum absolute error for $\beta_{i,k}(S_i)$ is 0.5% and the maximum absolute error for $\alpha_{i,k,\tilde{k}}$ is 0.9%.

For asymmetric instances, we report exact values for $\theta_{i,k}(S_i)$ (equal for each $k \in K$), and exact and approximate values for $\beta_{i,k}(S_i)$, $k \in K$. Errors in $A_{i,k}(S_i)$, $k \in K$, follow immediately from Equation (1) (for asymmetric instances, we do not study $\alpha_{i,k,\tilde{k}}(S_i)$). Table 3 shows us the results for instances as defined in Table 2. Again, we observe small deviations. Absolute errors are less than 1% for all $\beta_{i,k}(S_i)$, $k \in K$, in all instances.

In the above, we considered instances with main local warehouses only. That means that we specifically tested our approximation step described in §4.2 that decouples the mains. From the numerical results, we can conclude that this approximation step works well. It results in small errors in performance measures compared to exact results.

In the remainder of this subsection, we describe instances having both main and regular local warehouses. For instances as described in Table 4, we report results in Table 5. We report results of the exact and approximate evaluation for mains and regulars separately. Notice that in our approximate evaluation, $\beta_{i,j}(S_i)$ is determined exactly for the regulars $j \in J \setminus K$.

Table 4 Parameter settings for instances with mains and regulars, with $M_{i,k}$ in parts per year and with $t^{\text{reg}} = 0.04$ years, with $K = \{1, 2\}$, $\sigma(1) = (2)$ and $\sigma(2) = (1)$ for $|K| = 2$, with $k_j = 1$, $j \in J \setminus K$, for $|K| = 1$, and $J \setminus K = \{3, 4\}$, $k_3 = 1$ and $k_4 = 2$ for $|K| = 2$

Instance	$ K $	$ J - K $	$M_{i,k}$ ($k \in K$)	$M_{i,j}$ ($j \in J \setminus K$)	$S_{i,k}$ ($k \in K$)	$S_{i,j}$ ($j \in J \setminus K$)
1	1	1	0.5	0.5	1	1
2	1	1	1	1	1	1
3	1	1	5	5	1	1
4	1	1	10	10	1	1
5	1	1	50	50	1	1
6	1	1	5	10	1	1
7	1	1	10	5	1	1
8	1	1	5	10	1	2
9	1	1	5	10	2	1
10	1	1	10	5	2	1
11	1	2	0.5	0.5	1	1
12	1	2	1	1	1	1
13	1	2	5	5	1	1
14	1	2	10	10	1	1
15	1	2	50	50	1	1
16	1	2	5	10	1	1
17	1	2	10	5	1	1
18	1	2	5	10	1	2
19	1	2	5	10	2	1
20	1	2	10	5	2	1
21	2	2	0.5	0.5	1	1
22	2	2	1	1	1	1
23	2	2	5	5	1	1
24	2	2	10	10	1	1
25	2	2	50	50	1	1
26	2	2	5	10	1	1
27	2	2	10	5	1	1
28	2	2	5	10	1	2
29	2	2	5	10	2	1
30	2	2	10	5	2	1

As can be seen from Table 5, again results from the approximate evaluation are close to exact results, although the maximum absolute error is now 2%, obtained for $A_{i,j}(S_i)$, and thus also for $\theta_{i,j}(S_i)$, for the regulars in instance 6. This is higher than in instances with main local warehouses only. For the mains, the maximum absolute error is 1.1%, obtained in instance 16. Notice that for instances with one main (instances 1-20), it is clear that the decoupling of the regulars from the main is the one and only cause of errors, since for one main, no decoupling of mains takes place. For the instances with 2 mains (instances 21-30), both decoupling of the regulars from the mains and decoupling of the system of mains do take place.

For our numerical comparison, we used instances with a few local warehouses only. Computation times were small; we measured computation times in milliseconds, and they were mostly zero. However, for the exact evaluation, some instances required 10 milliseconds. So for those instances there is a large difference in computation time between the exact and the approximate evaluation. This difference will even be larger for instances with a larger number of local warehouses, and as $|J|$ becomes too large, exact evaluation will be impossible, due to the size of the state spaces.

In summary, we can conclude that our approximate evaluation is accurate and fast.

Table 5 Results for instances with mains and regulars

Instance	$\beta_{i,k}(S_i) (k \in K)$		$\beta_{i,j}(S_i) (j \in J \setminus K)$		$A_{i,k}(S_i) (k \in K)$		$A_{i,j}(S_i) (j \in J \setminus K)$		$\theta_{i,k}(S_i) (k \in K)$		$\theta_{i,j}(S_i) (j \in J \setminus K)$	
	exact	appr.	exact	appr.	exact	appr.	exact	appr.	exact	appr.	exact	appr.
1	0.980	0.980	0.980				0.019	0.019	0.0200	0.0200	0.0006	0.0004
2	0.960	0.960	0.962				0.036	0.037	0.0399	0.0399	0.0022	0.0015
3	0.812	0.811	0.833				0.127	0.135	0.1878	0.1892	0.0401	0.0315
4	0.665	0.660	0.714				0.172	0.189	0.3350	0.3396	0.1133	0.0970
5	0.238	0.231	0.333				0.143	0.154	0.7619	0.7692	0.5238	0.5128
6	0.767	0.761	0.714				0.198	0.217	0.2325	0.2391	0.0881	0.0683
7	0.699	0.698	0.833				0.109	0.116	0.3013	0.3023	0.0572	0.0504
8	0.820	0.819	0.946				0.040	0.044	0.1798	0.1814	0.0145	0.0098
9	0.962	0.964	0.714				0.267	0.275	0.0382	0.0362	0.0185	0.0103
10	0.938	0.939	0.833				0.153	0.156	0.0617	0.0615	0.0135	0.0102
11	0.980	0.980	0.980				0.019	0.019	0.0204	0.0204	0.0006	0.0004
12	0.959	0.959	0.962				0.036	0.037	0.0413	0.0413	0.0023	0.0016
13	0.792	0.789	0.833				0.124	0.132	0.2079	0.2105	0.0431	0.0351
14	0.622	0.614	0.714				0.162	0.175	0.3781	0.3860	0.1240	0.1103
15	0.184	0.176	0.333				0.112	0.118	0.8162	0.8235	0.5545	0.5490
16	0.711	0.700	0.714				0.184	0.200	0.2891	0.3000	0.1020	0.0857
17	0.684	0.682	0.833				0.107	0.114	0.3163	0.3182	0.0595	0.0530
18	0.807	0.804	0.946				0.039	0.043	0.1926	0.1957	0.0151	0.0106
19	0.937	0.940	0.714				0.259	0.268	0.0627	0.0604	0.0265	0.0173
20	0.931	0.931	0.833				0.152	0.155	0.0694	0.0691	0.0149	0.0115
21	0.980	0.980	0.980			0.020	0.020	0.020	0.0008	0.0008	0.0000	0.0000
22	0.959	0.959	0.962		0.038	0.038	0.038	0.038	0.0032	0.0032	0.0002	0.0001
23	0.784	0.783	0.833		0.147	0.148	0.152	0.155	0.0694	0.0691	0.0149	0.0115
24	0.595	0.592	0.714		0.199	0.201	0.217	0.227	0.2055	0.2068	0.0688	0.0591
25	0.149	0.145	0.333		0.113	0.112	0.164	0.171	0.7382	0.7435	0.5028	0.4957
26	0.724	0.720	0.714		0.168	0.172	0.245	0.255	0.1085	0.1082	0.0408	0.0309
27	0.640	0.639	0.833		0.193	0.193	0.135	0.139	0.1673	0.1675	0.0316	0.0279
28	0.793	0.793	0.946		0.143	0.144	0.049	0.051	0.0641	0.0637	0.0054	0.0034
29	0.958	0.962	0.714		0.037	0.035	0.284	0.285	0.0044	0.0035	0.0021	0.0010
30	0.931	0.932	0.833		0.059	0.058	0.164	0.165	0.0101	0.0099	0.0022	0.0017

5 Approximation

In the previous section, we have described how, for each SKU $i \in I$, a given policy, i.e., a choice for all base stock levels $S_{i,j}$, $j \in J$, can be evaluated in an approximate way. The current section deals with approximation, i.e., finding a feasible policy for Problem (P) with as low cost as possible. For this integer-programming problem with non-linear constraints, we provide a greedy heuristic, in which we use our approximate evaluation method to determine cost and waiting times of given policies.

Wong et al. (2005) have shown that with exact evaluation, the greedy algorithm followed by local search performs very well (they study instances with main local warehouses only and one customer group per local warehouse). We performed additional experiments with the greedy method and exact evaluation, but without local search. For instances with 50 and 100 SKU-s, and 2, 3, or 4 local warehouses, all of which are mains, we found gaps of at most 3.7% compared to a lower bound obtained by Dantzig-Wolfe decomposition / Lagrangian relaxation, and the size of this gap was decreasing in the number of SKU-s: the average gap for instances with 50 SKU-s was 1.00% and the average gap for instances with 100 SKU-s was 0.72%. This shows that the greedy method performs reasonably well.

Our greedy method builds up inventory for all SKU-s $i \in I$ in all local warehouses $j \in J$. It can be described in three steps. In the first (initialization) step, we set all base stock levels $S_{i,j} := 0$, $i \in I$, $j \in J$. In the second step, we increase base stock levels if and as long as it does not increase cost. We execute this step for each SKU $i \in I$ separately (since cost and waiting times do depend on S_i only). As long as for an SKU i an increase of a base stock level $S_{i,j}$ would lead to a cost decrease, we increase the base stock level that gives us the largest cost decrease by one. In the third step, as long as our current solution is not feasible, we iteratively increase the base stock level $S_{i,j}$, $i \in I$, $j \in J$, that provides us with the largest decrease in waiting time per unit cost increase.

For our algorithmic description, we introduce the following notation. Let e_j , $j \in J$, denote a row vector of size $|J|$ with the j -th element equal to 1 and all other elements equal to 0. Define $\Delta C(i, j) := C_i(S_i + e_j) - C_i(S_i)$, $i \in I$, $j \in J$, as the cost difference if the base stock level for SKU i at local warehouse j would be increased by one, at a given vector S_i . Furthermore, define the decrease in distance to the set of feasible policies if for SKU $i' \in I$ and local warehouse $j' \in J$ the base stock level $S_{i',j'}$ will be increased by one as

$$\Delta W(i', j') := \sum_{j \in J} \sum_{n \in N_j} \left[\sum_{i \in I} \frac{m_{i,n}}{M_n} W_{i,j}(S_i) - \hat{W}_n^{\text{obj}} \right]^+ - \sum_{j \in J} \sum_{n \in N_j} \left[\sum_{i \in I \setminus \{i'\}} \frac{m_{i,n}}{M_n} W_{i,j}(S_i) + \frac{m_{i',n}}{M_n} W_{i',j}(S_{i'} + e_{j'}) - \hat{W}_n^{\text{obj}} \right]^+,$$

with $[a]^+ := \max\{0, a\}$. Finally, define ratio $R(i, j)$, $i \in I$, $j \in J$, as $R(i, j) := \Delta W(i, j) / \Delta C(i, j)$.

In Algorithm 3, we describe our heuristic formally.

Algorithm 3

Step 1 Set $S_{i,j} := 0$, $i \in I$, $j \in J$.

Step 2 For each SKU $i \in I$:

Step 2-a Calculate $\Delta C(i, j)$, $j \in J$.

Step 2-b While $\min\{\Delta C(i, j)\} \leq 0$:

1. Determine $\hat{j}$ such that $\Delta C(i, \hat{j}) \leq \Delta C(i, j)$, $j \in J$.
2. Set $S_{i,\hat{j}} := S_{i,\hat{j}} + 1$.
3. Calculate $\Delta C(i, j)$, $j \in J$.

Step 3

Step 3-a Calculate $R(i, j)$, $i \in I$, $j \in J$.

Step 3-b While $\max\{R(i, j)\} > 0$:

1. Determine $\hat{i}$ and $\hat{j}$ such that $R(\hat{i}, \hat{j}) \geq R(i, j)$, $i \in I$, $j \in J$.
2. Set $S_{\hat{i},\hat{j}} := S_{\hat{i},\hat{j}} + 1$.
3. Calculate $R(i, j)$, $i \in I$, $j \in J$.

Computation time can be saved in the execution of the algorithm if only those (intermediate) results are updated that are affected by a certain change. E.g., waiting times $W_{i,j}(S_i)$, $j \in J$, only depend upon the base stock levels for SKU i , and are not subject to change if a base stock level for another SKU is increased.

Notice that an increase of a base stock level always leads to an increase in holding cost; see Equation (3). Step 2 of Algorithm 3 thus only increases base stock levels if the decrease in transportation cost exceeds the increase in holding cost. An increase of a base stock level always leads to a decrease in expected transportation cost and to a decrease in expected waiting time, as can be shown using a sample path argumentation.

6 Partial vs. full pooling

In this section we investigate numerically how partial pooling performs compared to the full pooling situation in terms of cost. We use the heuristic described in the previous section, where evaluation of given policies is done approximately, as described in Section 4.

We use a small data set with 5 local warehouses. Each local warehouse serves one group of machines. In our data set we have 50 SKU-s with fictitious cost data and equal demand rates per group. Cost prices for SKU-s 1, 2, ..., 50, are 2000, 4000, ..., 100000 Euro, respectively. The

Table 6 Results for data set with 50 SKU-s for cases with different $|K|$

$ K $	Total yearly cost (Euros)	Cost savings compared to no pooling (%)
0	2800766.21	
1	2188490.43	21.9
2	1929074.21	31.1
3	1886028.17	32.7
4	1819068.70	35.1
5	1818257.93	35.1

holding cost parameter for holding one item on stock for one day is calculated as $0.25/365$ times the SKU price. Demand rates per group per day for SKU-s 1, 2, ..., 50, are 0.0100, 0.0098, ..., 0.0002, respectively. Since each warehouse serves one group, the demand rate per day for an SKU i represents both $m_{i,n}$, $n \in N$, and $M_{i,j}$, $j \in J$.

Target waiting times are equal for all groups at the different local warehouses, and set to 0.10 day. All $t_{j,k}^{\text{lat}}$, $j \in J$, $k \in K$, are equal to 0.5 days. Further, all t_j^{em} , $j \in J$, are equal to 2 days and $t^{\text{reg}} = 14$ days. All $C_{j,k}^{\text{lat}}$, $j \in J$, $k \in K$ are equal to 500 Euro, and all C_j^{em} , $j \in J$, are equal to 1000 Euro.

For the data set described above, we study 6 different cases, where the number of main local warehouses $|K|$ varies from 0 to 5. In each case, the other $5 - |K|$ local warehouses are regulars. We use a cyclic lateral transshipment order, and assignments of regulars to mains are distributed as over the mains in an as equal as possible way. E.g., for the case with $|K|=2$ mains $\{1, 2\}$ and 3 regulars $\{3, 4, 5\}$, we then have $\sigma(1) = (2)$, $\sigma(2) = (1)$, $k_3 = k_5 = 1$, and $k_4 = 2$.

For these six cases, we report the total yearly cost for the solution obtained by the greedy algorithm in Table 6. Furthermore, we report the cost savings compared to the no pooling case with $|K| = 0$. Compared to the no pooling situation, the full pooling situation implies a cost saving of 35.1%. A very important observation here is that if we apply partial pooling with *one* main local warehouse(s) only, already $21.9 / 35.1 = 62.3\%$ of the full pooling benefits are obtained. This is an important observation. With two main local warehouses, we obtain already $31.1 / 35.1 = 88.7\%$ of the full pooling benefits. As we mentioned in Section 1, because of practical considerations it is often not desirable to let all local warehouses carry out lateral transshipment. Also, the operational cost may be too high to make all local warehouses main local warehouses. Our results imply that a major part of the possible benefits of using lateral transshipment is obtained by a low number of mains.

In addition to our results on the accuracy of our approximate evaluation method in Section 4, we make two remarks here that provide additional empirical evidence on the accuracy of our approximate evaluation method.

1. We evaluated the final solution obtained by the greedy algorithm in each of the six cases in Table 6 exactly. The differences in waiting time are small compared to the waiting times that we obtained by approximate evaluation: The waiting time values that we obtain deviate at most 1.52% from the exact waiting times.
2. We ran the greedy heuristic using exact evaluation instead of approximate evaluation. Although the final solution in some of the six cases differs slightly, the total cost is almost equal to the outcomes in Table 6: The values that we obtain for total cost deviate at most 2.02% from the total cost values obtained as final solution if in the greedy method exact evaluation is used.

The greedy algorithm with exact evaluation on average needed 163 seconds, where the results in Table 6, using our greedy algorithm with approximate evaluation, are obtained much quicker: in 30 milliseconds on average. In Section 4, we have shown that our approximate evaluation method obtains accurate results, i.e., results that are reasonably close to values as obtained by exact evaluation. The results obtained in this section confirm our believe that our approximate evaluation method is accurate. Furthermore, the gain in computation time is evident.

7 Case study: ASML

In this section, we perform a case study with data obtained from ASML. This data set constitutes data for all 19 local warehouses in the United States of America. In Sections 4 and 6, we have shown numerically that our approximate evaluation method as described in Section 4 is accurate. We will use this evaluation method here. The aim of this section is to show potential savings of pooling spare parts inventory at ASML. In line with the main result in Section 6, we will see that with partial pooling already a substantial part of the full pooling benefits can be obtained.

In this section, we first describe the ASML data set. Then, we compare the total cost for cases with different numbers of mains. Finally, we spend attention to savings compared to the old situation at ASML, i.e., the situation that lateral transshipment was not taken into account in the planning, but nevertheless used in practice.

The data set for the 19 local warehouses in the United States of America is as follows. We have 1451 SKU-s, where the price of the most expensive item is about 10^6 times the price of the cheapest item. Holding cost rates per year are 0.25 times the price. All $t_{j,k}^{\text{lat}}$, $j \in J$, $k \in K$, are equal to 0.5 days. Further, $t_j^{\text{em}} = 2$ days for all $j \in J$ and $t^{\text{reg}} = 14$ days. All $C_{j,k}^{\text{lat}}$, $j \in J$, $k \in K$ are equal to 500 Euro, and $C_j^{\text{em}} = 1000$ Euro for all $j \in J$.

In our case, in total 27 groups are present, and each of the 19 local warehouses serves one or two groups. Commonality is present in this data set, i.e., at local warehouses that serve two groups, these groups have some SKU-s in common. For each of the groups $n \in N$, the target waiting time

Table 7 Results for case study at ASML for cases with different $|K|$

$ K $	Total yearly cost (normalized)	Cost savings compared to no pooling (%)	Computation time (seconds)
0	100		71
4	51.82	48.18	57
19	49.81	50.19	132

Table 8 Results for case study at ASML: Comparison between old and new model at $|K| = 4$

Model	Target waiting time in days (avg)	Realized waiting time in days (avg)	Total yearly cost (normalized)
Old model	0.1500	0.1007	100
New model	0.1007	0.0975	68.51

$\hat{W}_n^{\text{obj}}$ is set to 0.15 days. Demand rates are low; the highest value for $m_{i,n}$ that occurs is 39 per year.

In the ASML data set, four local warehouses are identified by ASML as suitable candidates to be mains. A pre-specified order for lateral transshipment between these mains is given by ASML, as well as k_j for all regular local warehouses $j \in J \setminus K$.

To compare the total yearly cost at different numbers of main local warehouses, we study cases with $|K| = 0, 4$, and 19, respectively. At $|K| = 19$, as pre-specified lateral transshipment order we assume that first the four local warehouses are checked that are mains in the case with $|K| = 4$, and then all other local warehouses. Results are given in Table 7. The total yearly cost in the no pooling situation is normalized to 100. Table 7 shows that in the full pooling case more than 50% of the no pooling cost can be saved. In line with results obtained in the previous section, we observe that in the partial pooling situation with four mains, the majority of these potential savings can be obtained: $48.18 / 50.19 = 95.99\%$. Clearly, the additional value of having more main local warehouses is very small. Computation time results show that the algorithm is fast enough.

In Table 7, we have shown huge potential savings. In the last part of this section, we show another comparison.

Before implementation of our algorithms, ASML did not take lateral transshipment into account in the planning phase, i.e., the inventory in each local warehouse was planned separately. Nevertheless, in daily practice lateral transshipment was used. Given the parameter values for transportation times and cost at ASML, this meant that in practice, ASML obtained lower waiting times than it planned for, since actual performance increased because of the lateral transshipment option from the four main local warehouses. Suppose that these lower waiting times are set as target for the case with $|K| = 4$ in the new model, then the savings in yearly cost are 31.49%, as can be seen from Table 8.

For application in practice, our model is very appropriate since it incorporates lateral trans-

shipment in a convenient way. The inventory control method described in this paper has been implemented at ASML as part of a total planning concept for spare parts inventory control that is used since early 2005. By using our model in the total planning concept, ASML has reduced both waiting times and cost substantially.

8 Conclusion

We conclude this paper by summarizing our main results. First, we introduced a new form of partial pooling, which is general and matches with lateral transshipment procedures as observed in practice. Second, we developed an accurate and fast approximate evaluation method. That method has been used in an accurate and fast greedy heuristic for the optimization of base stock levels. Third, we showed that partial pooling performs very well compared to full pooling. If only a few of the local warehouses are allowed to provide lateral transshipment, then already a substantial part of the full pooling benefits is obtained. Finally, we executed a case study based on data of ASML, and showed that their spare parts planning could be significantly improved by our greedy algorithm. ASML implemented our algorithm and realized a substantial improvement in terms of waiting times and cost. In the past two years, we experimented with our algorithms at a few other companies, and also there they appeared to be applicable, and substantial potential savings in comparison to current procedures were identified; see Rijk (2007) and Van Sommeren (2007).

References

- Alfredsson, P., Verrijdt, J., 1999. Modeling emergency supply flexibility in a two-echelon inventory system, *Management Science* 45, 1416–1431.
- Archibald, T.W., 2007. Modelling replenishment and transshipment decisions in periodic review multilocation inventory systems, *Journal of the Operational Research Society* 58, 948–956.
- Archibald, T.W., Sassen, S.A., Thomas, L.C., 1997. An optimal policy for a two depot inventory problem with stock transfer, *Management Science* 43, 173–183.
- Axsäter, S., 1990. Modeling emergency lateral transshipments in inventory systems, *Management Science* 36, 1329–1338.
- Axsäter, S., 2003a. Evaluation of unidirectional lateral transshipments and substitutions in inventory systems, *European Journal of Operational Research* 149, 438–447.
- Axsäter, S., 2003b. A new decision rule for lateral transshipments in inventory systems, *Management Science* 49, 1168–1179.
- Evers, P.T., 2001. Heuristics for assessing emergency transshipments, *European Journal of Operational Research* 129, 311–316.
- Kranenburg, A.A., 2006. *Spare Parts Inventory Control under System Availability Constraints*, PhD dissertation, Technische Universiteit Eindhoven, Eindhoven, The Netherlands (available online: <http://alexandria.tue.nl/extra2/200612097.pdf>).

- Kukreja, A., C.P. Schmidt, D.M. Miller, 2001. Stocking decisions for low-usage items in a multilocation inventory system, *Management Science* 47, 1371–1383.
- Kutanoglu, E., 2008. Insights into inventory sharing in service parts logistics systems with time-based service levels, *Computers & Industrial Engineering* 54, 341–358.
- Lee, H.L., 1987. A multi-echelon inventory model for repairable items with emergency lateral transshipments, *Management Science* 33, 1302–1316.
- Minner, S., E.A. Silver, D.J. Robb, 2003. An improved heuristic for deciding on emergency transshipments, *European Journal of Operational Research* 148, 384–400.
- Rijk, P.J., 2007. *Multi-item, multi-location stock control with capacity constraints for the fieldstock of service parts at Océ*, MSc thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands.
- Ross, S.M., 1996. *Stochastic Processes*, Wiley, New York.
- Tagaras, G., Cohen, M.A., 1992. Pooling in two-location inventory systems with non-negligible replenishment lead times, *Management Science* 38, 1067–1083.
- Tijms, H.C., 2003. *A First Course in Stochastic Models*, Wiley, Chichester, England.
- Van Sommeren, R.J.M., 2007. *Spare parts inventory control for a single-echelon, multi-location model with lateral transshipments*, MSc thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands.
- Varga, R., 1962. *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey.
- Wong, H., Van Houtum, G.J., Cattrysse, D., Van Oudheusden, D., 2005. Simple, efficient heuristics for multi-item multi-location spare parts systems with lateral transshipments and waiting time constraints, *Journal of the Operational Research Society*, 1419–1430.
- Wong, H., Van Houtum, G.J., Cattrysse, D., Van Oudheusden, D., 2006. Multi-item spare parts systems with lateral transshipment and waiting time constraints, *European Journal of Operational Research* 171, 1071–1093.
- Xu, K., Evers, P.T., Fu, M.C., 2003. Estimating customer service in a two-location continuous review inventory model with emergency transshipments, *European Journal of Operational Research* 145, 569–584.
- Zhao, H., Deshpande, V., Ryan, J.K., 2005. Inventory sharing and rationing in decentralized dealer networks, *Management Science* 51, 531–547.