European Journal of Operational Research 276 (2019) 983–997

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing, Transportation and Logistics

The impact of an emergency warehouse in a two-echelon spare parts network

E. van Wingerden, T. Tan*, G. J. Van Houtum

School of Industrial Engineering, Eindhoven University of Technology, the Netherlands

ARTICLE INFO

Article history:
Received 22 November 2017
Accepted 28 January 2019
Available online 31 January 2019

Keywords:
Inventory
Heuristics
Spare parts
System-oriented service constraints

ABSTRACT

Because of high downtime costs of capital goods, it is becoming increasingly more important to have spare parts available as fast as possible. In order to ensure this, companies are resorting to the use of so-called lateral and emergency shipments in case the stock is not available at the closest local warehouse. As a result, managing the inventory of spare parts in a large network becomes increasingly more complex. In this paper we present a spare parts inventory model for a two-echelon spare parts network with lateral and emergency shipments and an emergency warehouse which is intended to allow for reserving stock nearby the central warehouse dedicated to emergency requests. Although any sequence is possible, the order is commonly fixed as it is based on the time and thus costs to get a part from each other location to the customer. We first present an accurate approximate procedure to evaluate such a complex network. Using simulation we find that this approximate evaluation procedure is accurate, especially for higher availability levels. We then look at optimized basestock levels, obtained via a smart enumeration procedure, to show the benefit of making use of this emergency warehouse. We find that savings up to 30% can be obtained via an emergency warehouse.

1. Introduction

This research is motivated by collaborating with ASML, an original equipment manufacturer (OEM) responsible for the production and maintenance of lithography machines all over the world. When maintaining a large installed base of capital goods, one of the key challenges is to keep total maintenance costs as low as possible. Especially for capital goods, the costs of downtime play a significant role in the total costs. Downtime costs can be even over 40% of the total costs of ownership (Öner, Kiesmüller, & Van Houtum, 2007). To reduce this downtime, a proper management of the spare parts inventory plays an important role as this allows for a quick replacement on site. This inventory can be managed by the users of the capital goods themselves, by the OEMs, or third parties. It is quite common that the OEM or third parties control the inventory as they can gather information about the demand rates from all customers to get more accurate forecasts, and they get an inventory pooling effect by sharing the available stock. When the OEM serves many customers, such as in the case of ASML, the OEM is responsible for the inventory control of multiple warehouses. The goal is to keep the inventory holding costs as low as possible while satisfying the service level agreements made with all customers.

In the spare parts inventory network of ASML, demands from the customers arrive at local warehouses nearby the customer. It is common to have a local warehouse nearby the customer as this allows to reduce the time it needs to get the parts where they are needed. Whenever there is no stock at that local warehouse and a spare part is requested, the spare part is retrieved from another location from where the spare part can be delivered as fast as possible. As the sort of shipment depends on the network structure there are many different possible sequences. Some companies may only have a few of these options whereas other companies can make use of each option to get the parts from other locations. One of these options is to provide the part from another local warehouse in the same region by using a lateral transshipment. Although the distance has a big impact on the time it takes to deliver the part, a shorter distance does not always guarantee a shorter time to deliver the part. Even if another local warehouse is located relatively close to the customer, a lateral transshipment may not always be the most suitable option. For instance, customs or available transport modes can play an important aspect in the time to get the part to the customer. Moreover, not every local warehouse is equipped with the necessary resources to allow for a fast emergency shipment or desire to ship away parts as they may be requested at their location later on. Therefore, another

* Corresponding author.
E-mail address: t.tan@tue.nl (T. Tan).

https://doi.org/10.1016/j.ejor.2019.01.068
0377-2217/

option to provide the part is to make use of a so-called emergency warehouse, which is a warehouse without its own external demand dedicated to ship spare parts as fast as possible when requested. This warehouse is located nearby the central warehouse, preferably at a strategic location such that spare parts can be put on transport by several different transport modes. An example of such a location is close to an airport. When a request for a spare part comes in, it can deliver the spare part faster than other local warehouses which may be located closer but are not as fast as the emergency warehouse due to the available transport modes and/or regulations regarding customs. However, depending on the network the emergency warehouse does not necessarily have to be the fastest option. Therefore, when demand arrives and the local warehouse does not have stock on hand, a preference list is followed, which states the order at which other warehouses are requested to deliver the spare part. This preference list is given and known, as in practice companies determine this preference list based on their priorities. In the case of ASML the time to deliver the part is the most important. As the sequence at which lateral or emergency shipments are applied may thus be different for every local warehouse we need an evaluation procedure which allows for this flexibility. Although other companies may have different networks, many international companies with warehouses in different regions are facing similar problems. Therefore, we model the problem in such a way that it applies to almost any company with similar characteristics.

Our goal is to minimize the total costs which include costs for waiting for a spare part as well as the costs for holding spare parts on stock, because spare parts can be very expensive. By taking the emergency warehouse into consideration we have more flexibility on where to stock parts and this allows for a reduction of the total costs.

The inventory control of spare parts has been a popular topic for research over the recent years. For a two-echelon network managed by a base stock policy without lateral and emergency shipments, Sherbrooke (1968) introduces the METRIC (Multi-Echelon Technique for Repairable Item Control) approach to estimate the expected number of backorders at the local warehouses. Sherbrooke (1968) approximates the replenishment lead times for the local warehouses by independent and deterministic lead times. This method is developed further by Graves (1985), where a negative binomial distribution is fitted to the first two moments of the pipeline stock. This approximation was shown to give more accurate results for the expected number of backorders than the results of the METRIC approach. Graves' method is generalized to multi-indenture systems later on by Rustenburg, van Houtum, and Zijm (2004). Muckstadt and Thomas (1980) provide an approximate evaluation procedure and focus on the optimization of the base stock levels by comparing centralized and decentralized decision making.

Besides the evaluation, Sherbrooke (1968) develops a heuristic optimization method to set the base stock levels to minimize the total holding costs while keeping the total expected number of backorders under the target for the whole system. For a similar problem, but without a constraint per local warehouse, Wong, Kranenburg, Van Houtum, and Cattrysse (2007) develop multiple heuristics. Andersson and Melchiors (2001) consider a two-echelon system where there are no backorders at the local warehouses, but demand is lost instead. They develop an accurate heuristic to determine a cost-effective base stock policy, based on the METRIC approximation. Muckstadt and Thomas (1980) extend the work of Sherbrooke (1968) by considering emergency shipments. They introduce a heuristic optimization of the base stock levels in order to compare the benefit of centralized decision making over decentralized decision making. Özkan, Van Houtum, and Serin (2015) develop a fast approximate evaluation procedure that was shown to outperform the approximation of Muckstadt and Thomas (1980). They use the observation that emergency requests are treated differently by the central warehouse than replenishment requests and improve the accuracy of their approximation by taking this into account explicitly. In this paper we make use of this idea presented at Özkan et al. (2015) for our approximation. Although all above methods consider two-echelon networks and thus take into account the impact of the stock of the central warehouse on the performance of the local warehouses, these methods do not take the use of lateral and/or emergency shipments into consideration.

For two-echelon networks where emergency shipments are sent only from the supplier to the local warehouses, and the central warehouse is faced by both replenishment requests as well as direct customer demand, Axsäter, Kleijn, and De Kok (2004) develops a heuristic method using a critical level at the warehouse to differentiate the demand streams based on their priorities. However, the use of lateral transshipments is not taken into consideration, thus this model is only suitable when local warehouses are located far apart from each other.

For a multi-echelon network with lateral transshipments, Axsäter (1990) develops an approximate evaluation based on the METRIC approach to estimate the replenishment lead time from the central warehouse to the local warehouses. The local warehouses are divided in groups, between which the use of lateral transshipments is possible. However, the use of emergency shipments is not possible.

Grahovac and Chakravarty (2001) consider a two-echelon network, where there is a possibility of lateral transshipments from other local warehouses as well as an emergency shipment from the central warehouse, similar to the network we consider in this paper. However, unlike many other papers with lateral and/or emergency shipments, the central warehouse is checked first whenever there is no stock at the local warehouse. Only if this is not possible, they look for a lateral transshipments at other local warehouses. This may be the case when distances between local warehouses are large or similar with respect to the distance to the central warehouse. Moreover, an emergency shipment from the supplier is not possible. Alfredsson and Verrijdt (1999) consider a two-echelon network with both lateral and emergency shipments. They assume the use of full pooling between local warehouses, thus representing the case that a lateral transshipment can be provided from each local warehouse to any other local warehouse, and if this is not possible they make use of an emergency shipment from the central warehouse or supplier. Making use of the full pooling assumption, they aggregate the total demand of the local warehouses in order to calculate the fraction of demand satisfied by emergency shipments by modeling the problem as a two-dimensional Markov process. They numerically compute the limiting distribution of the Markov process, which is shown to give accurate results, and even exact in the case that all local warehouses are identical. Unfortunately, this method is very time-consuming, even for small problem instances. Moreover, local warehouses in practice are generally not identical. In our paper, we not only tackle the problem of complexity but also allow for a more general network structure, including the use of an emergency warehouse and flexibility with respect to the preferred sequence at which demand is handled in case of a stockout at a local warehouse. The above mentioned papers are thus special cases of our problem.

Another stream of literature is the stream about single-echelon, multi-location networks with lateral and emergency shipments. Kutanoglu (2008) propose an iterative scheme for the evaluation in the case there is full pooling. As key performance indicator for the company, they use time-based service constraints. Kutanoglu and Mahajan (2009) look at a similar network, where they allow for prioritizing the warehouses for the lateral transshipments, and propose an implicit enumeration-based method to find the

minimum-cost stock levels. Wong, van Houtum, Cattrysse, and van Oudheusden (2005) develop a heuristic optimization method which is built on exact evaluations via Markov processes for a similar network. As the central warehouse is assumed to have ample stock, the emergency shipment can always be provided by the central warehouse. A similar system is studied by Kranenburg and Van Houtum (2009), where it is also possible to have a form of partial pooling. They introduce main warehouses, which are local warehouses which can provide other local warehouse by means of a lateral transshipment, and regular warehouses which do not provide lateral transshipments to other warehouses. They introduce an approximate evaluation procedure based on modeling lateral transshipments as Poisson overflow processes, similar to Axsäter (1990). In order to minimize the holding costs subject to a waiting time constraint, they develop an efficient greedy heuristic. This work has been implemented at ASML and has shown that just a few main warehouses can be sufficient to get most of the gains of lateral transshipments. For a more general system with lateral and emergency shipments with customers requesting spare parts, van Wijk, Adan, and Van Houtum (2012) develop an approximate heuristic where it is possible to define a preference list at which local warehouses are consulted for each customer. The method of Reijnen, Tan, and Van Houtum (2009) is a special case of the evaluation of van Wijk et al. (2012). The use of this preference list is similar to our case. For a complete overview of studies related to lateral transshipments, see Paterson, Kiesmüller, Teunter, and Glazebrook (2011). Although all of the above literature consider a certain form of lateral transshipments, they do not all allow for the same amount of flexibility, or take the central warehouse into consideration.

Finally, we look at the stream of literature that looks at the use of dedicated stock for emergency shipments, which has similarities to the use of an emergency warehouse. Axsäter, Howard, and Marklund (2013) considers a single-echelon, multi-location network where they introduce an emergency warehouse. This support warehouse is used to provide the part whenever a local warehouse runs out of stock, but does not directly face demand. However, the central warehouse is not taken into consideration. Howard, Marklund, Tan, and Reijnen (2015) look at a similar network, including a central warehouse with ample capacity. The customers can get the part from the support warehouse or the central warehouse in the case the local warehouse is out of stock when a request occurs. Making use of the pipeline information they achieve cost-efficient policies for requesting emergency shipments. van Wijk, Adan, and Van Houtum (2013) study a multi-location inventory problem with a so-called quick response warehouse and derive the optimal policy for when to make use of the quick response stock. Although these papers consider the use of an emergency warehouse, the use of their emergency warehouse is different to our paper. These emergency warehouses are located in the field and may not be able to serve all warehouses whereas our emergency warehouse is located nearby the central warehouse at a more central location and is always able to deliver any other warehouse. Moreover, in these papers the finite stock at the central warehouse and/or the use of lateral transshipment between local warehouses are not considered.

We contribute to the literature in the following ways:

- We introduce the use of an emergency warehouse, which is a warehouse dedicated for fast emergency shipments without having its own external demand, for a two-echelon network with lateral and emergency shipments. The problem setting we consider is motivated by ASML.
- We provide an accurate and efficient evaluation procedure for two-echelon networks that can make use of both lateral and emergency shipments while taking this emergency warehouse into consideration. We not only show that our evaluation procedure is accurate for the instances that we are mainly interested in, the evaluation is also sufficiently fast to be used in optimization procedures for problems of real-life size.
- Our evaluation procedure applies to a very general network structure. E.g., the network structures considered in Grahovac and Chakravarty (2001), Muckstadt and Thomas (1980), Özkan et al. (2015), and Alfredsson and Verrijdt (1999) are all special cases of our general structure. We allow any kind of sequence of lateral transshipments and/or an emergency shipment from the emergency warehouse.
- Using straightforward optimization, which is possible due to the speed of our approximate evaluation procedure, we show that cost savings of over 30% may occur compared to the case without an emergency warehouse (see Section 5).

The remainder of the paper is organized as follows. In Section 2, we give a formal description of the model. In Section 3 we present our approximate evaluation procedure. This procedure consists of two parts, a Local Evaluation Procedure and a Central Evaluation Procedure. We then perform a numerical study on a variety of different scenarios in Section 4 where we compare our evaluation procedure with simulation in order to measure its accuracy for different settings. Then in Section 5 we apply a smart enumeration procedure to obtain insights on the usefulness of the emergency in warehouse. Finally, we conclude our paper in Section 6.

2. Model description

We consider a single-item, two-echelon inventory model with a central warehouse (CW), an emergency warehouse, and one or more local warehouses (LW). Let $J = \{1, 2, \ldots, |J|\}$ be the set of local warehouses with $|J| \geq 1$. The index of the emergency warehouse is denoted by $|J| + 1$. The emergency warehouse, a warehouse without its own external demand dedicated to ship spare parts as fast as possible when requested, is located nearby the central warehouse at a strategic location such that spare parts can be put on transport by several different transport modes. It is also possible that the emergency warehouse is a separate warehouse within the central warehouse that can handle such an emergency request much faster compared to the central warehouse. Demand at each local warehouse is assumed to arrive according to a Poisson process with a constant rate μ_j, $j \in J$. The emergency warehouse (EW) does not face direct customer demand. The inventory in the network is controlled by a base stock policy. Hence, if a warehouse (a local warehouse, the emergency warehouse or the central warehouse) fulfills a demand, the inventory levels drop by one unit, and immediately a replenishment order is placed. In spare parts inventory systems, availability of (critical) spare parts is of crucial importance, as lack of spare parts might result in system down time. The cost associated with these down times (contractually or otherwise) typically overweigh the economies of scale cost benefit of batching in shipment or ordering (for the same SKU), as the risk of downtime increases while waiting to batch with lower stock. Consequently, in spare part inventory systems it is very common to use a base stock policy at all warehouses, both in theory and in practice (see e.g. Muckstadt, 2005; Sherbrooke, 2004; Van Houtum & Kranenburg, 2015). Also at ASML, such a base stock policy is applied. Let $K = J \cup \{|J| + 1\}$. We denote the base stock level as S_k, $k \in K \cup \{0\}$, where $\mathbf{S} = (S_0, S_1, S_2, \ldots, S_{|J|+1})$.

Whenever there is a request at a local warehouse and there is no stock on hand at that local warehouse, there is a fixed sequence at which other local warehouses and the emergency warehouse are checked to see if they can deliver the spare part. This can be any sequence, although it is commonly based upon the time it takes to get the part from that warehouse. Let v_j, $j \in J$, be the array

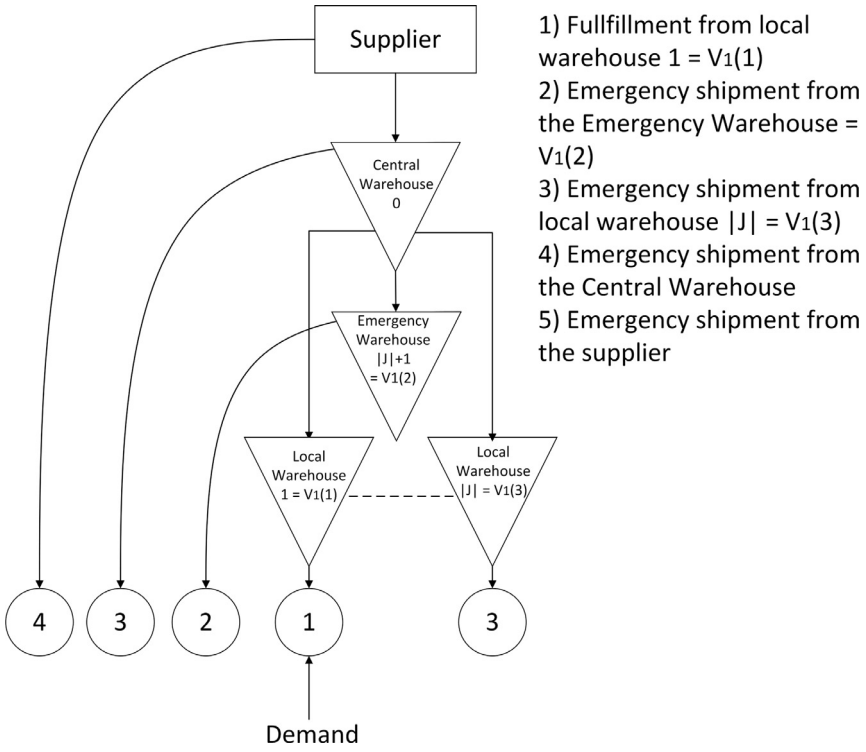

Fig. 1. Example order of demand fulfillment when demand arrives at warehouse 1, with $v_1 = (1, |J|+1, |J|)$.

consisting of the sequence at which warehouses are checked for an emergency shipment (for convenience of notation both lateral transshipments and emergency shipments are called emergency shipments from here on), with $v_j(i)$ the ith warehouse in the array $(i = 1, 2, \ldots, p_j)$, with $p_j \geq 1$ representing the maximum number of warehouses that is checked in case of a stockout. The warehouses in this sequence can consist of both local warehouses as well as the emergency warehouse. If the warehouse where the demand arrives does not have stock, it will check warehouses $v_j(1)$ up and to $v_j(p_j)$. When none of the local warehouses or the emergency warehouse have stock, the central warehouse, denoted by index 0, is checked for an emergency shipment. If the central warehouse has stock on hand it will supply the spare part and immediately request a replenishment at the supplier. When there is also no stock at the central warehouse, an emergency shipment will be requested from the supplier, denoted by index -1. A graphical representation of the order at which the possible shipment options are used, given the example sequence $v_1 = (1, |J|+1, |J|)$ when demand arrives at local warehouse 1 is presented in Fig. 1.

We assume that the supplier has ample stock, and the supply lead time from the supplier to the central warehouse is denoted by $t_0 > 0$. The deterministic transportation time for a replenishment from the central warehouse to local warehouse j is denoted by t_j. Notice that because the emergency warehouse is in close proximity of the central warehouse, the replenishment lead time from the central warehouse to the emergency warehouse is assumed to be zero ($t_{|J|+1} = 0$).

Let us introduce the set $Q = K \cup \{-1, 0\}$. As we are mainly interested in the expected time each customer has to wait for a spare part, and the costs involved to provide a part, which depends on where the spare part is delivered from, we need to calculate $\theta_{q,j}$, the fraction of demand of local warehouse j, $j \in J$, that is served from location q, $q \in Q$. Note that by definition for each local warehouse $j \in J$ it holds that $\sum_{q \in Q} \theta_{q,j} = 1$.

Whenever warehouse $q \in Q$ is used to fulfill the demand which arrives at local warehouse $j \in J$, there are also costs involved for the emergency shipment and waiting time from location q to j, denoted by $c^{em}_{q,j}$. These costs consist of two different cost factors. First of all, this includes the costs for fast transportation and handling, denoted by $c^{ship}_{q,j}$. Secondly, and most important, this includes the costs for the additional standstill as a result of not having the part available at local warehouse j. For every hour the machine is not operating there are costs involved denoted by c^{down}_j. The time it takes in hours to obtain a part from warehouse q to j is denoted by $t^{em}_{q,j}$. As a result, the total costs involved when a part is delivered from warehouse q to warehouse j is calculated as follows:

$$c^{em}_{q,j} = c^{ship}_{q,j} + c^{down}_j t^{em}_{q,j}$$

Whenever the demand can be fulfilled by the local warehouse where the demand has arrived, we assume $c^{em}_{j,j} = 0$. As such, $c^{em}_{q,j}$ denotes the additional emergency costs compared to fulfilling the demand from the local warehouse directly. Next to the emergency costs, for each demand that is sent to the customer from warehouse q there are costs to replenish warehouse q denoted by c^{rep}_q. In the case of the supplier, these costs are zero. In the case of the central warehouse, it involves costs for replenishing the central warehouse. In the case of a local or emergency warehouse, it involves both the costs of a replenishment to the central warehouse as well as the local or emergency warehouse. The total costs are then denoted as follows:

$$\hat{C}(\mathbf{S}) = \sum_{k \in K \cup \{0\}} h S_k + \sum_{j \in J} \mu_j \left(\sum_{q \in Q} \theta_{q,j} (c^{em}_{q,j} + c^{rep}_q) \right),$$

where h represents the holding costs rate for each spare part kept on stock. Note that for the holding costs we assume that the costs of the pipeline inventory are included, which are the costs over the time between placing the order at the supplier and receiving the actual part at the central warehouse and the time it takes to replenish a local warehouse from the central warehouse. We apply the same holding cost rate for every warehouse as value added

from transport is minor. If holding cost rates differ between warehouses, it is possible to make the holding cost rates location dependent by changing h to h_k in the formula above. The analysis in the remainder of the paper would largely remain the same under such location dependent rates. The only change would occur in the optimization procedure in Section 5.1, where the lower bound cost function should be changed.

If there would be no lateral or emergency shipments, and thus all demand is satisfied by the local warehouses where demand arrived directly, one would observe the following replenishment costs:

$$C^{rep} = \sum_{j \in J} \mu_j c_j^{rep}$$

These costs C^{rep} form a constant factor, i.e. they are independent of the basestock policy. Hence, we can subtract them from the total costs function $\hat{C}(\mathbf{S})$ and then obtain the following new costs function $C(\mathbf{S})$:

$$\begin{aligned} C(\mathbf{S}) &= \hat{C}(\mathbf{S}) - C^{rep} \\ &= \sum_{k \in K \cup \{0\}} hS_k + \sum_{j \in J} \mu_j \left(\sum_{q \in Q} \theta_{q,j} (c_{q,j}^{em} + c_q^{rep}) \right) \\ &\quad - \sum_{j \in J} \mu_j c_j^{rep} \sum_{q \in Q} \theta_{q,j} \\ &= \sum_{k \in K \cup \{0\}} hS_k + \sum_{j \in J} \mu_j \left(\sum_{q \in Q} c_{q,j} \theta_{q,j} \right), \end{aligned}$$

where $c_{q,j} = c_q^{rep} - c_j^{rep} + c_{q,j}^{em}$. The overall goal is to minimize the total costs $C(\mathbf{S})$.

3. Evaluation procedure

In this section we explain our approximate evaluation procedure to obtain the systems performance for given base stock levels. We avoid numerical evaluations of Markov processes as in the procedure of Alfredsson and Verrijdt (1999), because that leads to too long computation times. Instead, we decouple all locations and incorporate their dependencies in an appropriate way. A fast performance evaluation is then obtained by iteratively analyzing individual locations and updating of the dependencies.

As the central warehouse has a finite stock, whenever the central warehouse runs out of stock and a replenishment request comes in, the central warehouse has to wait for a replenishment from the supplier before it can send the part to the local warehouse or emergency warehouse. By adding the average waiting time for a part at the central warehouse, W_0, to the fixed transport time, we estimate the average replenishment lead time from the central warehouse as follows:

$$t_k^{reg} = t_k + W_0, \qquad k \in K. \tag{1}$$

To apply our approximate evaluation algorithm, we decouple the network into two parts as shown in Fig. 2 making use of Eq. (1). By decoupling the problem into two separate parts, which do still depend on each other, we obtain two smaller evaluation procedures which can be evaluated efficiently.

The Local Evaluation Procedure, which is further explained in Section 3.1, consists of the evaluation of the local warehouses and the emergency warehouse. For the evaluation we consider the emergency warehouse as another warehouse without an own external demand process, which can always be checked for an emergency shipment by all other local warehouses. Given the average replenishment lead time from the central warehouse, t_k^{reg}, we evaluate the performance of the local warehouses and the emergency warehouse. The output of this evaluation procedure is used as an input for the Central Evaluation Procedure.

The Central Evaluation Procedure, which is further explained in Section 3.2, consists of the evaluation of the central warehouse. For this procedure we are interested in the average waiting time, which is used as an input for the Local Evaluation Procedure and depends on the ratio between replenishment and emergency requests. As these two evaluation procedures depend on the output of the other evaluation procedure we propose an iterative procedure in Section 3.3, where we combine the two evaluation procedures into a single evaluation procedure.

3.1. Local Evaluation Procedure

For the Local Evaluation Procedure we adapt the evaluation procedure of Reijnen et al. (2009) to include the use of the emergency warehouse. The approach of Reijnen et al. (2009) is shown to be efficient and usable for large problem instances (see also van Wijk et al., 2012).

Demand arrives at the local warehouses only with rate $M_{j,j} = \mu_j$, $j \in J$. The local warehouses try to deliver the part from stock and if not possible, they try to obtain the part from another warehouse. Firstly, warehouse j is checked. If warehouse j does not have stock this results in so-called overflow demand from warehouse j to warehouse $v_j(1)$. Let us now introduce the following notation which we use in our Local Evaluation Procedure:

- $M_{k,j}$: The demand rate for warehouse $k, k \in K$, originating from local warehouse j, $j \in J$.
- M_k: Total demand rate for warehouse $k, k \in K$, including overflow demand ($= \sum_{j \in J} M_{k,j}$).
- β_k: The fraction of total demand that warehouse $k \in K$ is able to directly deliver from stock.

For the evaluation, we assume that the warehouses are independent and that β_j is known. Then an approximation of the average demand rate for warehouse $v_j(1)$, originating from warehouse j, is as follows:

$$M_{v_j(1),j} = (1 - \beta_j) M_{j,j}. \tag{2}$$

Using the fraction of demand that is satisfied by warehouse $v_j(1)$, $\beta_{v_j(1)}$, the demand warehouse $v_j(2)$ faces, originating from warehouse j, is on average $M_{v_j(2),j} = (1 - \beta_{v_j(1)}) M_{v_j(1),j}$. For $2 \le i \le p_j$ these demand rates can be expressed as follows:

$$M_{v_j(i),j} = (1 - \beta_{v_j(i-1)}) M_{v_j(i-1),j} \tag{3}$$

As the emergency warehouse can be checked by any local warehouse, the emergency warehouse is included in every sequence v_j. See Fig. 3 for a graphical representation of the overflow of demand where the emergency warehouse is at the end of the sequence. We approximate that each overflow demand stream follows a Poisson process. Under this assumption the total demand at warehouse k also follows a Poisson process with rate M_k. Under this approximation we can evaluate each location independently of each other. As we have emergency shipments, the number of parts in the pipeline is at most S_k for local warehouse k. As a result the behavior of the number of parts on order for location k is as the number of jobs in an $M|G|c|c$ queue with $c = S_k$ parallel servers. This queue is also called an Erlang loss system (see e.g. Karush, 1957). The fill rate can thus be calculated using known results from the Erlang loss system. The Erlang loss probability is given by:

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

where ρ represents the offered load. The fill rate can thus be calculated as 1 minus the fraction of time that all servers are occupied, with S_k representing the number of servers, and $t_k^{reg} M_k$, the

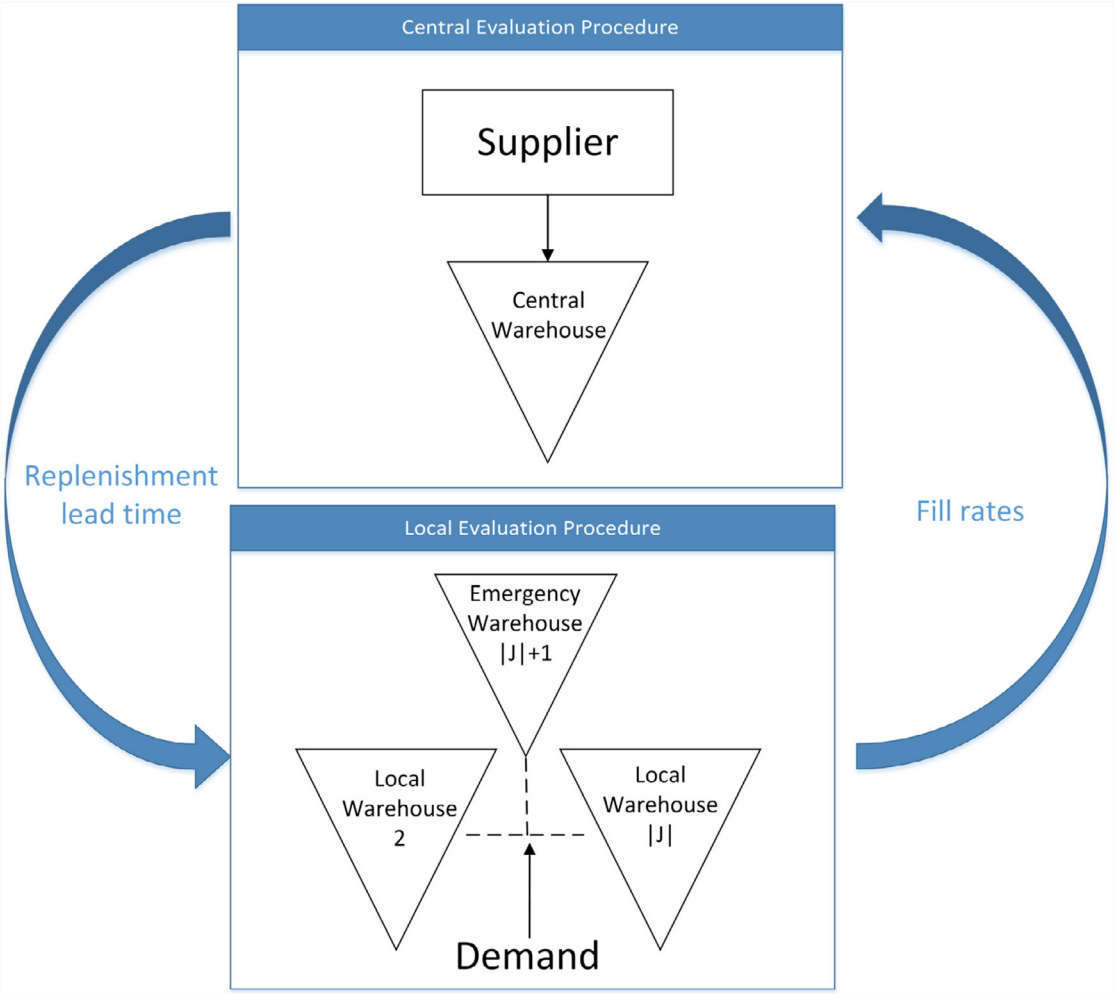

Fig. 2. Network decoupling and interaction.

demand during the leadtime arriving at warehouse k, representing the offered load:

$$\beta_k = 1 - L(S_k, t_k^{\text{reg}} M_k) \qquad k \in K, \tag{4}$$

Using the fill rate and the demand rate, we calculate the fraction of demand for warehouse j that is served by warehouse k:

$$\theta_{k,j} = \frac{\beta_k M_{k,j}}{\mu_j} \qquad k \in K, j \in J.$$

To determine the values of M_k, $M_{k,j}$ and β_k we use an iterative procedure. For the first iteration we assume there is no overflow demand at all, thus $M_{j,j} = \mu_j$ and $M_{v_j(i),j} = 0$ for $1 \le i \le p_j$, $j \in J$. For each warehouse $k \in K$ we determine M_k, and calculate β_k using Eq. (4), where t_k^{reg} is given. After this step we make an improved estimation of the demand using Eq. (3). These steps are repeated until M_k does not change more than ϵ, with ϵ small. The Local Evaluation Procedure is described in Algorithm 1. Although we are not able to prove that this algorithm always converges, in our experiments it always did.

Algorithm 1 Local Evaluation Procedure.

Step 1. Initialization

$M_{j,j} := \mu_j$ $\quad \forall j \in J$

$M_{k,j} := 0$ $\quad \forall k \in K, j \in J, k \neq j$

$M_k := \sum_{j \in J} M_{k,j}$ $\quad \forall k \in K$

$\beta_k := 1 - L(S_k, t_k^{\text{reg}} M_k)$ $\quad \forall k \in K$

Step 2. Compute β_k, and M_k until each M_k does not change more than ϵ

$M_{v_j(i),j} = (1 - \beta_{v_j(i-1)}) M_{v_j(i-1),j}$ $\quad \forall j \in J$ and $1 \le i \le p_j$

$M_k = \sum_{j \in J} M_{k,j}$ $\quad \forall k \in K$

$\beta_k = 1 - L(S_k, t_k^{\text{reg}} M_k)$ $\quad \forall k \in K$

Step 3. Calculate the performance

$\theta_{k,j} = \frac{\beta_k M_{k,j}}{\mu_j}$ $\quad k \in K, j \in J.$

3.2. Central Evaluation Procedure

In this section we introduce the Central Evaluation Procedure which evaluates the performance of the central warehouse. We estimate the average waiting time at the central warehouse, W_0, which we need in order to estimate the average replenishment lead times to the local warehouses and the emergency warehouse. We split this evaluation procedure into two cases; the case where we make use of the emergency warehouse ($S_{|J|+1} > 0$), and the case where we do not make use of the emergency warehouse ($S_{|J|+1} = 0$). If we do not make use of the emergency warehouse, i.e. if we have a base stock level of zero at the emergency warehouse, then two different demand streams arrive at the central warehouse; replenishment requests and emergency orders of demand that could not be satisfied by local warehouses. If we make use of the emergency warehouse, then the central warehouse

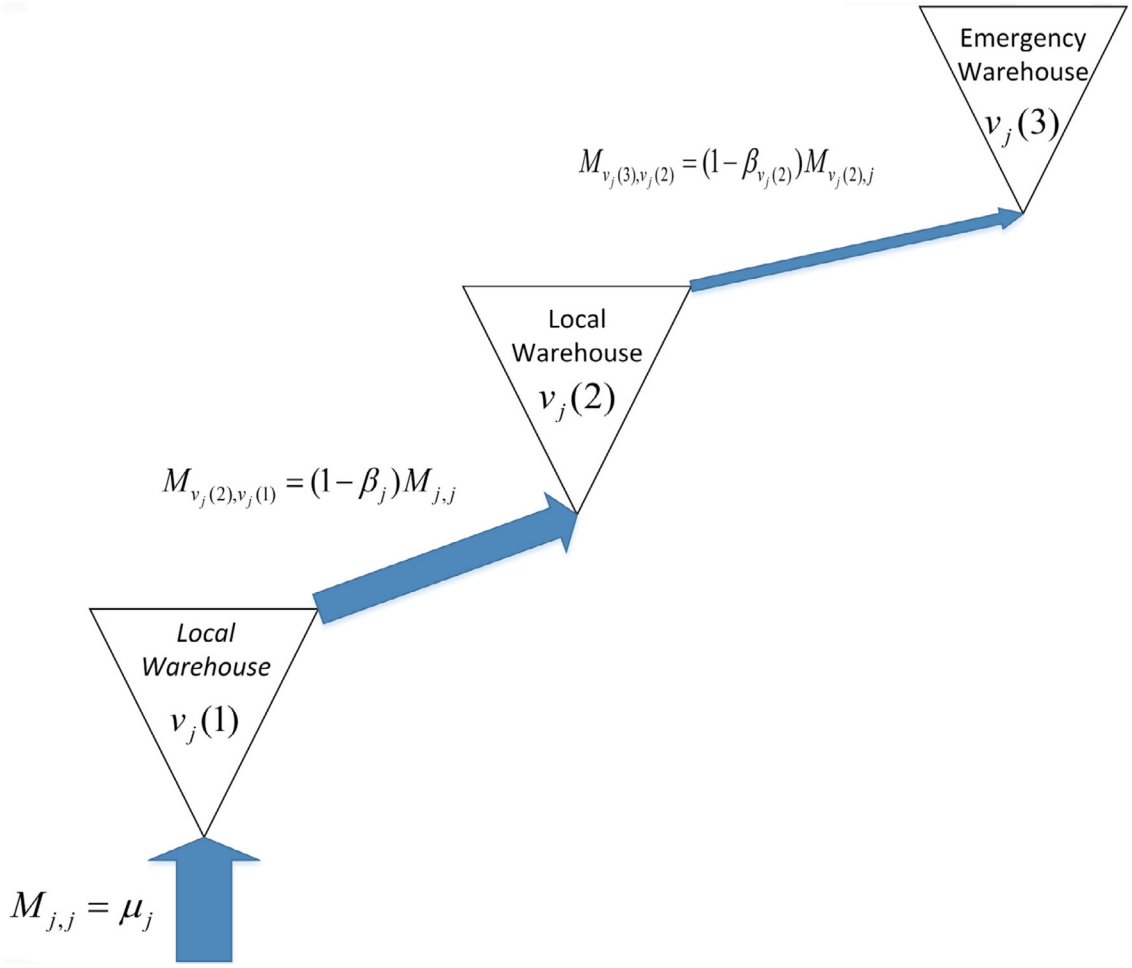

Fig. 3. Illustration of overflow demand.

only sees replenishment requests. This is due to the fact that the transportation time to the emergency warehouse is zero, hence; if the emergency warehouse does not have stock, neither does the central warehouse. In both cases we assume we know the fill rates of each local warehouse and the emergency warehouse as well as the demand rates for each warehouse.

3.2.1. Central evaluation when the emergency warehouse is not used

In the case we do not have an emergency warehouse, or do not decide to make use of it, we make use of the idea of Özkan et al. (2015). However, as we have lateral transshipments between the local warehouses in our model, we have to adapt the overall method of Özkan et al. (2015).

As long as the central warehouse has stock on hand, all demand that arrives at the local warehouses is satisfied by the central warehouse, either by a replenishment or by providing an emergency shipment. Thus the demand rate of the central warehouse as long as it has stock on hand is

$$M_0 = \sum_{j \in J} \mu_j. \tag{5}$$

When the central warehouse does not have stock on hand the replenishment requests still arrive at the central warehouse, as these requests are backordered. As the emergency requests will not be backordered but supplied by the supplier directly instead, the central warehouse does not observe these requests. As a result, the total demand rate for the central warehouse when it does not have stock on hand is equal to the demand rate for replenishment requests, which is the rate at which demand is satisfied by the local warehouses:

$$M_0' = \sum_{k \in K} M_k \beta_k. \tag{6}$$

We model the inventory level at the central warehouse as a birth-death process, i.e. a continuous-time Markov process with states $x \leq S_0$ representing the inventory level. For our approximate evaluation procedure we assume that the leadtime t_0 of the supplier is exponential with the same mean, i.e. by exponential times with rate $\mu_0 = \frac{1}{t_0}$. Moreover, the possible states for the central warehouse are truncated by setting the maximum number of backorders to $\bar{S} = \sum_{k \in K} S_k$ as there can never be more backorders at the central warehouse. The steady state probabilities then satisfy the following equations (Özkan et al., 2015):

$$\pi_x = \begin{cases} \frac{M_0'}{(S_0 - x)\mu_0} \pi_{x+1}, & -\bar{S} \leq x < 0 \\ \frac{M_0}{(S_0 - x)\mu_0} \pi_{x+1}, & 0 \leq x < S_0 \end{cases}$$

By expressing everything as a function of π_{S_0}, π_{S_0} follows from the normalization. The mean number of backorders and the mean waiting time (using Little's law) are then:

$$B_0 = \sum_{x=-\bar{S}}^{-1} (-x)\pi_x$$

$$W_0 = \frac{B_0}{M_0'}$$

Given this mean waiting time, we update the replenishment lead time for each warehouse $k \in K$ by Eq. (1). The fraction of demand that is fulfilled by an emergency shipment from the central warehouse is approximated as follows:

$$\theta_{0,j} = \frac{\beta_0(1-\beta_{v_j(p_j)})M_{v_j(p_j),j}}{\mu_j},$$

where we look at the fraction of demand that is not delivered by the last warehouse in the sequence of warehouse j, and where β_0 represents the fill rate at the central warehouse, which is calculated as follows:

$$\beta_0 = \sum_{x=1}^{S_0} \pi_x.$$

As demand is satisfied by either one of the local warehouses, the emergency warehouse, the central warehouse or the supplier through the use of an emergency shipment, we know that $\sum_{i=1}^{p_j} \theta_{v_j(i),j} + \theta_{0,j} + \theta_{-1,j} = 1$, for all $j \in J$. We then get the following expression for the fraction of demand that is satisfied by an emergency shipment from the supplier:

$$\theta_{-1,j} = 1 - \theta_{0,j} - \sum_{i=1}^{p_j} \theta_{v_j(i),j} \tag{7}$$

3.2.2. Central evaluation when using the emergency warehouse

When we make use of the emergency warehouse, there is a small change due to the relation between the central warehouse and the emergency warehouse. Because the transport time from the central warehouse to the emergency warehouse is zero, we know that as long as the central warehouse has stock, so does the emergency warehouse. As a result we know that as long as the central warehouse has stock on hand, all demand arriving at the local warehouses is provided by either the local warehouses and/or the emergency warehouse. Therefore, we know the demand rate for the central warehouse as long as it has stock on hand follows Eq. (5).

Whenever the central warehouse does not have stock on hand, only the replenishment requests arrive at the central warehouse as the emergency requests are directly delivered by the supplier. In this case the demand rate is equal to Eq. (6).

We thus make use of the same equations as in Section 3.2.1 to calculate the expected waiting time, W_0.

Because we know that whenever the emergency warehouse does not have stock on hand, neither does the central warehouse, we know that the central warehouse never delivers the part by an emergency request to the local warehouses. As a result the fraction of demand satisfied by the central warehouse by an emergency shipment is as follows:

$$\theta_{0,j} = 0.$$

After this step we use Eq. (7) to estimate the fraction of total demand delivered by the supplier.

3.2.3. Central evaluation algorithm

A complete overview of the Central Evaluation Procedure is described in Algorithm 2. In this procedure, the calculation of $\theta_{0,j}$, and $\theta_{-1,j}$ is omitted, because that calculation is only needed at the end of the overall evaluation procedure.

Note that in the special case that no stock is kept at the central warehouse, thus $S_0 = 0$, the central warehouse places an order at the supplier to replenish the warehouse as soon as a replenishment request arrives. As the supplier lead time, t_0, is fixed, the replenishment lead time is then as follows:

$$t_k^{\text{reg}} = t_k + t_0,$$

Algorithm 2 Central Evaluation Procedure.

Step 1. Initialization
$W_0 := 0$
$M_0 := \sum_{j \in J} \mu_j$
$\bar{S} := \sum_{k \in K} S_k$

Step 2. Compute W_0
$M_0' := \sum_{k \in K} M_k \beta_k$
$\pi_x := \begin{cases} \frac{M_0'}{(S_0 - x)\mu_0}, & -\bar{S} \le x < 0 \\ \frac{M_0}{(S_0 - x)\mu_0}, & 0 \le x < S_0 \end{cases}$
$B_0 := \sum_{x=-\bar{S}}^{-1} (-x)\pi_x$
$W_0 := \frac{B_0}{M_0'}$

which is then no longer an approximation but exact. Therefore we use this expression in the case we keep no stock at the central warehouse.

3.3. Overall evaluation procedure

In this section we describe the overall evaluation procedure where we combine Algorithms 1 and 2 into a single algorithm. As each of the two evaluation procedures relies on the output of the other procedure, we use an iterative procedure. We start by selecting a starting value of zero for the waiting time at the central warehouse and initializing the necessary variables. Then we apply the Local Evaluation Procedure, followed by applying the central warehouse procedure. Then if the difference between the waiting time of central warehouse compared to the previous run, or the initialized value, is smaller than ϵ with ϵ small, the iterative procedure is finished and we only need to determine the fraction of demand that is satisfied by each location. Note that although we are not able to prove that this method converges, in our experiments it always did.

The overall evaluation procedure is described in Algorithm 3.

Algorithm 3 Overall Evaluation Procedure.

Step 1. Initialization
$W_0 := 0$
$M_0 := \sum_{j \in J} \mu_j$
$t_k^{\text{reg}} = t_k + W_0 \quad \forall k \in K$
$M_{j,j} := \mu_j \quad \forall j \in J$
$M_{k,j} := 0 \quad \forall k \in K, j \in J, k \ne j$
$M_k := \sum_{j \in J} M_{k,j} \quad \forall k \in K$
$\beta_k := 1 - L(S_k, t_k^{\text{reg}} M_k) \quad \forall k \in K$

Step 2. Apply Step 2 of the Local Evaluation Procedure described in Algorithm 1

Step 3. Apply Step 2 of the Central Evaluation Procedure described in Algorithm 2

Step 4. Repeat Steps 2 and 3 until W_0 does not change more than ϵ

Step 5. Finalization
$\theta_{k,j} := \frac{\beta_k M_{k,j}}{\mu_j} \quad \forall k \in K, j \in J$
If $S_{|J|+1} = 0$: $\theta_{0,j} := \frac{\beta_0(1-\beta_{v_j(p_j)})M_{v_j(p_j),j}}{\mu_j} \quad \forall j \in J$
If $S_{|J|+1} > 0$: $\theta_{0,j} := 0 \quad \forall j \in J$
$\theta_{-1,j} := 1 - \theta_{0,j} - \sum_{i=1}^{p_j} \theta_{v_j(i),j} \quad \forall j \in J$

4. Numerical results

In this section we test our approximate evaluation procedure on its accuracy by comparing the results with results obtained by

(a) No-No, No-Part, No-Full network structure

(b) Part-No, Part-Full network structure

(c) Full-no network structure

Fig. 4. Network structures which require a different sequence. Local warehouses are denoted by triangles and the central and emergency warehouse are denoted together by the stars symbol.

Table 1
Detailed order for emergency shipments for $|J| = 8$.

LW	No-Part	No-Full	Part-No	Part-Full	Full-No
1	{1,9,2,3,4}	{1,9,2,3,4,5,6,7,8}	{1,2,3,4,9}	{1,2,3,4,9,5,6,7,8}	{1,2,3,4,5,6,7,8,9}
2	{2,9,3,4,1}	{2,9,3,4,1,6,7,8,5}	{2,3,4,1,9}	{2,3,4,1,9,6,7,8,5}	{2,3,4,1,6,7,8,5,9}
3	{3,9,4,1,2}	{3,9,4,1,2,7,8,5,6}	{3,4,1,2,9}	{3,4,1,2,9,7,8,5,6}	{3,4,1,2,7,8,5,6,9}
4	{4,9,1,2,3}	{4,9,1,2,3,8,5,6,7}	{4,1,2,3,9}	{4,1,2,3,9,8,5,6,7}	{4,1,2,3,8,5,6,7,9}
5	{5,9,6,7,8}	{5,9,6,7,8,1,2,3,4}	{5,6,7,8,9}	{5,6,7,8,9,1,2,3,4}	{5,6,7,8,1,2,3,4,9}
6	{6,9,7,8,5}	{6,9,7,8,5,2,3,4,1}	{6,7,8,5,9}	{6,7,8,5,9,2,3,4,1}	{6,7,8,5,2,3,4,1,9}
7	{7,9,8,5,6}	{7,9,8,5,6,3,4,1,2}	{7,8,5,6,9}	{7,8,5,6,9,3,4,1,2}	{7,8,5,6,3,4,1,2,9}
8	{8,9,5,6,7}	{8,9,5,6,7,4,1,2,3}	{8,5,6,7,9}	{8,5,6,7,9,4,1,2,3}	{8,5,6,7,4,1,2,3,9}

simulation, which we consider to be the exact results. We consider 72 different instances where we look at networks with either 8 or 16 local warehouses. We consider two different demand rates per warehouse, a demand rate of 0.02 per week representing slow moving parts, and a demand rate of 0.1 per week representing faster moving parts. Moreover, we consider five different sequences at which the demand is fulfilled in case the local warehouse where demand arrives does not have stock on hand. First of all, we consider the sequence "No-No", where we do not have any lateral transshipments and the emergency warehouse is requested as a first resort when the local warehouse does not have any stock on hand. The second sequence we consider is the "No-Partial" sequence. In this case, first the emergency warehouse is checked and if this warehouse is not able to deliver, all local warehouses in the same region are checked for a lateral transshipment. For the sequence "No-Full" we also first look at the emergency warehouse and then for a lateral transshipment. In this case all other local warehouses can be checked for a lateral transshipment. Fig. 4(a) represents the network structure with locations for which these two sequences makes sense.

Next, we have the network structure as presented in Fig. 4(b). One possible sequence for such a structure is the "Partial-No", where we first check for lateral transshipment at the local warehouses in the same region and then consider the emergency warehouse. For the sequence "Partial-Full" we have a similar order, except that we also look at the other remaining local warehouses after checking the emergency warehouse as this might still be faster than getting a part from the supplier. Our last sequence "Full-No" represents the case where we first look at all local warehouses for a lateral transshipment and at the emergency warehouse only in case none of these local warehouses are able to deliver the part. Fig. 4(c) gives a possible network structure for which this sequence is appropriate. Table 1 gives the detailed order for these sequences for each warehouse in the case of 8 local warehouse, consisting of two regions. Note that for 16 warehouses the order is similar, as we then have two regions of 8 local warehouses instead of 4. Moreover, we consider different combinations of base stock levels, which altogether gives us a total of 72 instances.

The simulation software used to obtain the exact results is Omnet++. For each instance we use a warm-up period of 50,000 years, followed by a period of 2,500,000 years. For each instance, we did 25 replications. The results for the instances with 8 local warehouses are presented in Table 2, and the results for the instances with 16 local warehouses are presented in Table 3, together with a 99% confidence interval. Because the emergency warehouse can be in between lateral transshipments, we split up the total fraction of demand satisfied by lateral transshipments into the fraction of lateral transshipments before consulting the emergency warehouse, α_j^1, and the fraction of lateral transshipments after consulting the emergency warehouse, α_j^2. Let e be the index denoting the emergency warehouse in the sequence v_j, we then calculate α_j^1 and α_j^2 as follows:

$$\alpha_j^1 = \sum_{x=2}^{e-1} \theta_{v_j(x),j}$$

$$\alpha_j^2 = \sum_{x=e+1}^{p_j} \theta_{v_j(x),j}$$

Note that in some sequences we do not have lateral transshipments or only before or after consulting the emergency warehouse. As α_j^2 or α_j^1 then do not exist, this is denoted by "N/A". From the tables we can observe that in general our evaluation procedure is accurate as the differences with the simulation results are very limited. Especially if we look at the higher fill-rates and/or higher demand rates, the evaluation procedure turns out to be more accurate. This is as expected, because our assumption that overflow demand also follows a Poisson process is more reasonable for these instances. However, it should be noted that especially for the case that the fill rate of the local warehouses are lower, in the range of 0–60% say, differences tend to become larger although this should not be a big issue as generally the service provided to the customers is on the higher range with expected fill rates over 90% in general due to the high downtime costs. Moreover, the computation time per instances is in the range of 1–70 milliseconds. For instances with more local warehouses, as well as for instances with higher demand rates, the run time increases. The number of iterations needed between the central and Local Evaluation Procedure has a big influence on the run time. We found that the number

Table 2
Results for $|J| = 8$, $t_0 = 20, t_j = 3$ for all $j \in J, t_{|J|+1} = 0$.

Inst.	N	μ_j	S_0	S_j	$S_{\|J\|+1}$	Sequence	β_j		α_j^1		α_j^2		$\theta_{\|J\|+1,j}$		$\theta_{-1,j}$	
							Appr.	Sim.	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.	Appr.	Sim
1	8	0.02	1	1	1	No-No	0.7511	0.7571 ± 0.0009	N/A	N/A	N/A	N/A	0.1616	0.1394 ± 0.0005	0.0873	0.1035 ± 0.0007
2			1	1	1	No-Partial	0.7198	0.7187 ± 0.0008	N/A	N/A	0.1055	0.1195 ± 0.0006	0.1723	0.1489 ± 0.0005	0.0024	0.0129 ± 0.0002
3			1	1	1	No-Full	0.7190	0.7145 ± 0.0007	N/A	N/A	0.1085	0.1330 ± 0.0006	0.1725	0.1498 ± 0.0005	0.0000	0.0027 ± 0.0001
4			1	1	1	Partial-No	0.6648	0.6738 ± 0.0007	0.3226	0.2921 ± 0.0006	N/A	N/A	0.0123	0.0241 ± 0.0003	0.0003	0.0100 ± 0.0002
5			1	1	1	Partial-Full	0.6646	0.6709 ± 0.0007	0.3228	0.2923 ± 0.0006	0.0003	0.0092 ± 0.0002	0.0123	0.0246 ± 0.0003	0.0000	0.0030 ± 0.0001
6			1	1	1	Full-No	0.6604	0.6633 ± 0.0008	0.3394	0.3281 ± 0.0008	N/A	N/A	0.0002	0.0055 ± 0.0001	0.0000	0.0031 ± 0.0001
7			2	1	1	No-No	0.8118	0.8223 ± 0.0007	N/A	N/A	N/A	N/A	0.1495	0.1189 ± 0.0005	0.0387	0.0588 ± 0.0004
8			2	1	1	No-Partial	0.8006	0.8011 ± 0.0007	N/A	0N/A	0.0434	0.0679 ± 0.0004	0.1557	0.1257 ± 0.0004	0.0003	0.0053 ± 0.0001
9			2	1	1	No-Full	0.8005	0.7993 ± 0.0007	N/A	N/A	0.0438	0.0737 ± 0.0005	0.1557	0.1261 ± 0.0005	0.0000	0.0009 ± 0.0001
10			2	1	1	Partial-No	0.7644	0.7689 ± 0.0006	0.2325	0.2150 ± 0.0005	N/A	N/A	0.0031	0.0121 ± 0.0002	0.0000	0.0040 ± 0.0001
11			2	1	1	Partial-Full	0.7644	0.7677 ± 0.0007	0.2325	0.2152 ± 0.0006	0.0000	0.0037 ± 0.0001	0.0031	0.0123 ± 0.0002	0.0000	0.0011 ± 0.0001
12			2	1	1	Full-No	0.7637	0.7641 ± 0.0007	0.2363	0.2328 ± 0.0007	N/A	N/A	0.0000	0.0020 ± 0.0001	0.0000	0.0011 ± 0.0001
13			1	2	1	No-No	0.9587	0.9568 ± 0.0005	N/A	N/A	N/A	N/A	0.0378	0.0341 ± 0.0003	0.0035	0.0091 ± 0.0002
14			1	2	1	No-Partial	0.9584	0.9557 ± 0.0005	N/A	N/A	0.0036	0.0097 ± 0.0002	0.0380	0.0346 ± 0.0003	0.0000	0.0000 ± 0.0000
15			1	2	1	No-Full	0.9584	0.9557 ± 0.0005	N/A	N/A	0.0036	0.0097 ± 0.0002	0.0380	0.0346 ± 0.0003	0.0000	0.0000 ± 0.0000
16			1	2	1	Partial-No	0.9554	0.9516 ± 0.0005	0.0446	0.0483 ± 0.0005	N/A	N/A	0.0000	0.0001 ± 0.0000	0.0000	0.0000 ± 0.0000
17			1	2	1	Partial-Full	0.9554	0.9516 ± 0.0005	0.0446	0.0483 ± 0.0005	0.0000	0.0000 ± 0.0000	0.0000	0.0001 ± 0.0000	0.0000	0.0000 ± 0.0000
18			1	2	1	Full-No	0.9554	0.9516 ± 0.0005	0.0446	0.0484 ± 0.0005	N/A	N/A	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000
19		0.1	4	3	1	No-No	0.8295	0.8290 ± 0.0003	N/A	N/A	N/A	N/A	0.0576	0.0502 ± 0.0001	0.1129	0.1208 ± 0.0003
20			4	3	1	No-Partial	0.7573	0.7452 ± 0.0004	N/A	N/A	0.1780	0.1825 ± 0.0003	0.0621	0.0556 ± 0.0002	0.0026	0.0167 ± 0.0001
21			4	3	1	No-Full	0.7554	0.7326 ± 0.0005	N/A	N/A	0.1824	0.2063 ± 0.0004	0.0622	0.0560 ± 0.0002	0.0000	0.0051 ± 0.0001
22			4	3	1	Partial-No	0.7206	0.7125 ± 0.0005	0.2733	0.2611 ± 0.0004	N/A	N/A	0.0057	0.0117 ± 0.0001	0.0004	0.0147 ± 0.0001
23			4	3	1	Partial-Full	0.7203	0.7022 ± 0.0005	0.2736	0.2638 ± 0.0004	0.0004	0.0162 ± 0.0001	0.0057	0.0124 ± 0.0001	0.0000	0.0054 ± 0.0001
24			4	3	1	Full-No	0.7167	0.6932 ± 0.0005	0.2833	0.2980 ± 0.0005	N/A	N/A	0.0000	0.0033 ± 0.0000	0.0000	0.0055 ± 0.0001
25			8	3	2	No-No	0.8977	0.8964 ± 0.0003	N/A	N/A	N/A	N/A	0.0867	0.0707 ± 0.0002	0.0156	0.0328 ± 0.0002
26			8	3	2	No-Partial	0.8916	0.8782 ± 0.0003	N/A	N/A	0.0182	0.0445 ± 0.0002	0.0902	0.0756 ± 0.0002	0.0000	0.0017 ± 0.0000
27			8	3	2	No-Full	0.8916	0.8769 ± 0.0003	N/A	N/A	0.0182	0.0470 ± 0.0002	0.0902	0.0757 ± 0.0002	0.0000	0.0030 ± 0.0000
28			8	3	2	Partial-No	0.8650	0.8461 ± 0.0003	0.1347	0.1477 ± 0.0003	N/A	N/A	0.0003	0.0050 ± 0.0001	0.0000	0.0012 ± 0.0000
29			8	3	2	Partial-Full	0.8650	0.8453 ± 0.0003	0.1347	0.1481 ± 0.0003	0.0000	0.0012 ± 0.0000	0.0003	0.0051 ± 0.0001	0.0000	0.0003 ± 0.0000
30			8	3	2	Full-No	0.8649	0.8421 ± 0.0004	0.1351	0.1568 ± 0.0004	N/A	N/A	0.0000	0.0008 ± 0.0000	0.0000	0.0003 ± 0.0000
31			8	2	1	No-No	0.7659	0.7685 ± 0.0003	N/A	N/A	N/A	N/A	0.0906	0.0767 ± 0.0002	0.1435	0.1548 ± 0.0003
32			8	2	1	No-Partial	0.6361	0.6422 ± 0.0004	N/A	N/A	0.2560	0.2258 ± 0.0004	0.0949	0.0838 ± 0.0002	0.0130	0.0482 ± 0.0002
33			8	2	1	No-Full	0.6222	0.6110 ± 0.0004	N/A	N/A	0.2824	0.2778 ± 0.0004	0.0951	0.0845 ± 0.0002	0.0003	0.0267 ± 0.0002
34			8	2	1	Partial-No	0.5781	0.6012 ± 0.0003	0.3902	0.3259 ± 0.0004	N/A	N/A	0.0254	0.0267 ± 0.0001	0.0063	0.0462 ± 0.0002
35			8	2	1	Partial-Full	0.5718	0.5749 ± 0.0004	0.3946	0.3242 ± 0.0003	0.0068	0.0438 ± 0.0002	0.0266	0.0294 ± 0.0001	0.0002	0.0277 ± 0.0002
36			8	2	1	Full-No	0.5523	0.5603 ± 0.0004	0.4461	0.3986 ± 0.0003	N/A	N/A	0.0016	0.0127 ± 0.0001	0.0000	0.0284 ± 0.0002

Table 3
Results for $|J| = 16$, $t_0 = 20, t_j = 3$ for all $j \in J, t_{|J|+1} = 0$.

Inst.	N	μ_j	S_0	S_j	$S_{\|J\|+1}$	Sequence	β_j		α_j^1		α_j^2		$\theta_{\|J\|+1,j}$		$\theta_{-1,j}$	
							Appr.	Sim.	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.	Appr.	Sim
37	16	0.02	1	1	1	No-No	0.7216	0.7224 ± 0.0008	N/A	N/A	N/A	N/A	0.1136	0.1073 ± 0.0004	0.1648	0.1703 ± 0.0006
38			1	1	1	No-Partial	0.6506	0.6487 ± 0.0007	N/A	N/A	0.2283	0.2323 ± 0.0008	0.1210	0.1156 ± 0.0006	0.0001	0.0034 ± 0.0001
39			1	1	1	No-Full	0.6505	0.6472 ± 0.0007	N/A	N/A	0.2285	0.2367 ± 0.0008	0.1210	0.1158 ± 0.0005	0.0000	0.0003 ± 0.0000
40			1	1	1	Partial-No	0.6027	0.6053 ± 0.0006	0.3967	0.3866 ± 0.0006	N/A	N/A	0.0006	0.0058 ± 0.0001	0.0000	0.0023 ± 0.0001
41			1	1	1	Partial-Full	0.6027	0.6044 ± 0.0006	0.3967	0.3868 ± 0.0006	0.0000	0.0025 ± 0.0001	0.0006	0.0059 ± 0.0001	0.0000	0.0004 ± 0.0000
42			1	1	1	Full-No	0.6024	0.6022 ± 0.0007	0.3976	0.3968 ± 0.0007	N/A	N/A	0.0000	0.0006 ± 0.0000	0.0000	0.0000 ± 0.0000
43			2	1	1	No-No	0.7573	0.7597 ± 0.0007	N/A	N/A	N/A	N/A	0.1207	0.1093 ± 0.0005	0.1220	0.1310 ± 0.0006
44			2	1	1	No-Partial	0.7071	0.7035 ± 0.0008	N/A	N/A	0.1651	0.1770 ± 0.0008	0.1278	0.1178 ± 0.0005	0.0000	0.0017 ± 0.0001
45			2	1	1	No-Full	0.7070	0.7027 ± 0.0008	N/A	N/A	0.1652	0.1793 ± 0.0008	0.1278	0.1178 ± 0.0005	0.0000	0.0001 ± 0.0000
46			2	1	1	Partial-No	0.6642	0.6651 ± 0.0008	0.3356	0.3306 ± 0.0008	N/A	N/A	0.0002	0.0032 ± 0.0001	0.0000	0.0011 ± 0.0000
47			2	1	1	Partial-Full	0.6642	0.6647 ± 0.0008	0.3356	0.3308 ± 0.0008	0.0000	0.0011 ± 0.0000	0.0002	0.0033 ± 0.0001	0.0000	0.0002 ± 0.0000
48			2	1	1	Full-No	0.6641	0.6635 ± 0.0008	0.3359	0.3362 ± 0.0008	N/A	N/A	0.0000	0.0002 ± 0.0000	0.0000	0.0001 ± 0.0000
49			1	2	1	No-No	0.9466	0.9458 ± 0.0003	N/A	N/A	N/A	N/A	0.0414	0.0373 ± 0.0002	0.0119	0.0169 ± 0.0002
50			1	2	1	No-Partial	0.9453	0.9437 ± 0.0003	N/A	N/A	0.0125	0.0183 ± 0.0002	0.0422	0.0380 ± 0.0002	0.0000	0.0000 ± 0.0000
51			1	2	1	No-Full	0.9453	0.9437 ± 0.0003	N/A	N/A	0.0125	0.0183 ± 0.0002	0.0422	0.0380 ± 0.0002	0.0000	0.0000 ± 0.0000
52			1	2	1	Partial-No	0.9409	0.9388 ± 0.0002	0.0591	0.0612 ± 0.0002	N/A	N/A	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000
53			1	2	1	Partial-Full	0.9409	0.9388 ± 0.0002	0.0591	0.0612 ± 0.0002	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000
54			1	2	1	Full-No	0.9409	0.9388 ± 0.0002	0.0591	0.0612 ± 0.0002	N/A	N/A	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000
55		0.1	4	3	1	No-No	0.7901	0.7898 ± 0.0003	N/A	N/A	N/A	N/A	0.0314	0.0303 ± 0.0002	0.1785	0.1799 ± 0.0002
56			4	3	1	No-Partial	0.6452	0.6243 ± 0.0003	N/A	N/A	0.3221	0.3340 ± 0.0004	0.0325	0.0324 ± 0.0001	0.0002	0.0093 ± 0.0001
57			4	3	1	No-Full	0.6450	0.6150 ± 0.0003	N/A	N/A	0.3225	0.3502 ± 0.0003	0.0325	0.0325 ± 0.0001	0.0000	0.0023 ± 0.0000
58			4	3	1	Partial-No	0.6172	0.5953 ± 0.0003	0.3823	0.3914 ± 0.0003	N/A	N/A	0.0005	0.0052 ± 0.0000	0.0000	0.0080 ± 0.0000
59			4	3	1	Partial-Full	0.6172	0.5881 ± 0.0003	0.3823	0.3941 ± 0.0004	0.0000	0.0099 ± 0.0001	0.0005	0.0055 ± 0.0001	0.0000	0.0024 ± 0.0001
60			4	3	1	Full-No	0.6168	0.5825 ± 0.0003	0.3832	0.4140 ± 0.0003	N/A	N/A	0.0000	0.0011 ± 0.0000	0.0000	0.0024 ± 0.0001
61			8	3	1	No-No	0.8315	0.8309 ± 0.0004	N/A	N/A	N/A	N/A	0.0348	0.0328 ± 0.0001	0.1336	0.1363 ± 0.0003
62			8	3	1	No-Partial	0.7387	0.7178 ± 0.0004	N/A	N/A	0.2254	0.2442 ± 0.0003	0.0359	0.0351 ± 0.0002	0.0000	0.0028 ± 0.0001
63			8	3	1	No-Full	0.7387	0.7150 ± 0.0004	N/A	N/A	0.2254	0.2494 ± 0.0004	0.0359	0.0352 ± 0.0001	0.0000	0.0004 ± 0.0000
64			8	3	1	Partial-No	0.7159	0.6917 ± 0.0003	0.2841	0.3039 ± 0.0003	N/A	N/A	0.0000	0.0022 ± 0.0001	0.0000	0.0022 ± 0.0001
65			8	3	1	Partial-Full	0.7159	0.6898 ± 0.0003	0.2841	0.305 ± 0.0003	0.0000	0.0026 ± 0.0001	0.0000	0.0022 ± 0.0000	0.0000	0.0004 ± 0.0000
66			8	3	1	Full-No	0.7159	0.6875 ± 0.0003	0.2841	0.3117 ± 0.0003	N/A	N/A	0.0000	0.0003 ± 0.0000	0.0000	0.0005 ± 0.0000
67			8	4	1	No-No	0.9273	0.9265 ± 0.0002	N/A	N/A	N/A	N/A	0.0268	0.0234 ± 0.0001	0.0459	0.0501 ± 0.0002
68			8	4	1	No-Partial	0.9135	0.9063 ± 0.0003	N/A	N/A	0.0584	0.0687 ± 0.0002	0.0281	0.0249 ± 0.0001	0.0000	0.0000 ± 0.0000
69			8	4	1	No-Full	0.9135	0.9064 ± 0.0002	N/A	N/A	0.0584	0.0687 ± 0.0002	0.0281	0.0249 ± 0.0001	0.0000	0.0000 ± 0.0000
70			8	4	1	Partial-No	0.9062	0.8971 ± 0.0003	0.0938	0.1029 ± 0.0003	N/A	N/A	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000
71			8	4	1	Partial-Full	0.9062	0.8970 ± 0.0003	0.0938	0.1029 ± 0.0003	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000
72			8	4	1	Full-No	0.9062	0.8970 ± 0.0003	0.0938	0.1030 ± 0.0003	N/A	N/A	0.0000	0.0000 ± 0.0000	0.0000	0.0000 ± 0.0000

Table 4
Asymmetric instance description.

Warehouse j	μ_j	S_j
1	0.02	1
2	0.02	2
3	0.1	4
4	0.04	2
5	0.08	3
6	0.02	4
7	0.02	1
8	0.04	3

of iterations needed is between 1 and 16. We can conclude that our evaluation procedure is accurate and fast enough if we would apply it on real life symmetric instances. We also consider asymmetric scenarios, where the demand rate and base stock levels differ per local warehouse, to analyse whether our approximation is still accurate. Table 4 gives an overview of the demand rates and base stock levels, considering a scenario with 8 warehouses. We use the same sequences as provided in Table 1 and the same lead times as we did for the symmetric instances. The results are shown in Table 5. The results show the average absolute error, measured by the difference between the outcome of the approximation and simulation. Based on these results, we can conclude that the evaluation method is still accurate.

5. The benefit of an emergency warehouse

In this section we show the benefit of having an emergency warehouse. We first calculate the optimal base stock levels. Then we calculate the optimal base stock level for the same instance, but where we are not allowed to have a positive stock at the emergency warehouse, thereby representing the case we do not have an emergency warehouse. We then compare the costs of these two solutions to get insights about the benefit of having an emergency warehouse.

We describe our optimization procedure to determine the base stock levels in Section 5.1. Because an incremental heuristic like the greedy heuristic does not necessarily lead to the optimal solution, we resort to a smart enumeration procedure. In Section 5.2 we get insights about the benefit of having an emergency warehouse by comparing the results for different instances.

5.1. Optimization of the base stock levels

In this section we describe the smart enumeration procedure that we use to set the optimal base stock levels for both cases where we do and do not have an emergency warehouse, assuming the evaluation procedure is correct. The problem we want to solve is as follows:

$$\min C(\mathbf{S}) = \sum_{k\in K\cup\{0\}} hS_k + \sum_{j\in J}\mu_j\left(\sum_{q\in Q} c_{q,j}\theta_{q,j}\right)$$

$$S_k \in \mathbb{N}_0 \quad \text{for all} \quad k \in K\cup 0,$$

where $c_{j,j}=0$. Let us define $C^*(l)$ as the lowest costs over all feasible solutions with a total stock of exactly l, thus $\sum_{k=0}^{|J|+1} S_k = l$. We then start with $l=0$, and increase l to find the lowest overall costs. To determine at which value of l we can be certain that an increase would never lead to a better solution anymore we first define a lower bound on the total costs. Let us introduce $\hat{\beta}_j(S_j) = 1 - L(S_j, t_j\mu_j)$ as the upper bound on the fill-rate as we ignore waiting time from the central warehouse and do not include overflow demand from other local warehouses due to lateral transshipments. Let us first introduce the set $Z(l)$ that consists of all possible solutions that satisfy $\sum_{k\in K\cup 0} S_k = l$. Moreover, note that $\sum_{q\in Q\setminus\{j\}}\theta_{q,j} = (1-\beta_j(\mathbf{S}))$ by definition. We define $C_{LB}(l)$, a lower bound on the total costs as:

$$C_{LB}(l) = hl + \min_{\mathbf{S}\in Z(l)}\left(\sum_{j\in J}\mu_j(1-\hat{\beta}_j(S_j))\min_{q\in Q\setminus\{j\}}\{c_{q,j}\}\right)$$

$$C_{LB}(l) \le hl + \min_{\mathbf{S}\in Z(l)}\left(\sum_{j\in J}\mu_j(1-\beta_j(S_j))\min_{q\in Q\setminus\{j\}}\{c_{q,j}\}\right)$$

$$\le hl + \min_{\mathbf{S}\in Z(l)}\left(\sum_{j\in J}\mu_j\left(\sum_{q\in Q} c_{q,j}\theta_{q,j}\right)\right) = C^*(l)$$

Using known results that $L(c,\rho)$ is decreasing as a function of c, see Karush (1957), we can show that $C_{LB}(l)$ is convex and we are able to find the optimal values for S_k given l. Moreover, let u denote the value that minimizes $C_{LB}(u)$. Having this lower bound on the total costs, we know that we can stop the enumeration as soon as $C^* \le C_{LB}(l+1)$, and $l \ge u$, where C^* is the best solution found so far during the enumeration procedure, and which is set at a very high value initially.

We then obtain the following smart enumeration procedure to obtain the optimal base stock levels, $\mathbf{S}^*$:

In this procedure, solutions are evaluated by the approximate evaluation procedure. Hence we cannot guarantee that we do find optimal solutions, although it is likely the result are close-to-optimal as our approximate evaluation procedure is shown to be accurate.

Algorithm 4 Smart enumeration procedure.

Step 1. Let $l=0, C^*=\infty$

Step 2. For each $\mathbf{S}$ with $\sum_{j=0}^{|J|+1} S_k = l$, compute $C(\mathbf{S})$. If $C(\mathbf{S}) \le C^*, C^* = C(\mathbf{S})$, and $\mathbf{S}^* = \mathbf{S}$

Step 3. If $C^* \le C_{LB}(l)$, and $l > u$, stop. Else, $l = l+1$ and go to Step 2

5.2. Comparison

We apply the enumeration procedure on a number of different network structures and demand and cost settings. The network structure and corresponding costs are inspired by the service network of ASML, a service network where downtime costs are considerable. The different values for the demand and holding cost values cover a wide range of values found in real life service networks. For all cases we assume the central warehouse is located in the Netherlands. First we look at a network structure that is suitable for the "No-No", "No-Part" and "No-Full" sequence; see Fig. 4(a). Table 6 describes the cost structure for this network. The costs are motivated by the time it takes for a warehouse from that location to the other location and multiplied by a fictive cost per hour for having downtime of the system. Note that the costs can be scaled to any number and still obtain the same solution as long as the cost ratio stays the same. For this network local warehouses are located relatively close to the central and thus emergency warehouse. We assume two warehouses to be located in England and two warehouses to be located in Germany. For such a structure it is interesting to first go for an emergency shipment instead of a lateral. For the lateral transshipments we can distinguish two cases, the case where we only retrieve a part from the warehouse of the same country, and the case where we also allow lateral transshipments from Germany to England and the other way around. Note that the difference in costs between the emergency warehouse and central warehouse is mainly due to

Table 5
Results for asymmetric instances, $|J| = 8$, $t_0 = 20, t_j = 3$ for all $j \in J, t_{|J|+1} = 0$.

Inst.	N	Sequence	β_j Average abs. error (%)	α_j^1 Average abs. error	α_j^2 Average abs. error	$\theta_{|J|+1,j}$ Average abs. error (%)	$\theta_{-1,j}$ Average abs. error (%)
73	8	No-No	0.42	N/A	N/A	1.01	1.16
74		No-Part	0.62	N/A	1.50%	0.96	0.02
75		No-Full	0.63	N/A	1.53%	0.96	0.00
76		Part-No	0.76	0.71%	N/A	0.05	0.01
77		Part-Full	0.76	0.71%	0.02%	0.05	0.00
78		Full-No	0.78	0.78%	N/A	0.00	0.00

Table 6
Network structure for No-Part and No-Full sequence.

$c_{q,j}$	Local warehouse 1	Local warehouse 2	Local warehouse 3	Local warehouse 4
Local warehouse 1	0	3000	4500	3750
Local warehouse 2	3000	0	3750	4500
Local warehouse 3	4500	3750	0	3000
Local warehouse 4	3750	4500	3000	0
Emergency warehouse	2250	2250	2250	2250
Central warehouse	5250	5250	5250	5250
Supplier	27,000	27,000	27,000	27,000

Table 7
Network structure for No-Part and No-Full sequence.

$c_{q,j}$	Local warehouse 1	Local warehouse 2	Local warehouse 3	Local warehouse 4
Local warehouse 1	0	4500	18,000	18,000
Local warehouse 2	4500	0	18,000	18,000
Local warehouse 3	18,000	18,000	0	3000
Local warehouse 4	18,000	18,000	3000	0
Emergency warehouse	7500	7500	10,500	10,500
Central Warehouse	10,500	10,500	13,500	13,500
Supplier	27,000	27,000	27,000	27,000

the fact that parts are received faster when send from the emergency warehouse compared to the central warehouse. This can have multiple reasons, for example, parts can be picked faster, there may be more opportunities to send the part, or agreements have been made with customs for parts in this warehouse, as well as the strategic location of the emergency warehouse (i.e. nearby an airport). The cost structure in the second network (see Fig. 4(b)), representing the "Part-No" and "Part-Full"sequence is described by Table 7. For this network local warehouses are further away from the central warehouse, which makes lateral transshipments more interesting than an emergency shipment. However, the local warehouses are divided over two regions for which the distance is further than the distance to the central and emergency warehouse, thus it is not interesting to first go for a lateral transshipment at all the local warehouses. For this case we assume there are two local warehouses located in the East coast of the USA and two local warehouses in Taiwan. Our final network structure represents the "Full-No" sequence and is not further presented in this paper because there were no instances for which the emergency warehouse gave any benefit. When we have a group of relatively closely related warehouses and the central warehouse is further away it makes sense to first go for a lateral transshipment at any of the other warehouses before going for an emergency shipment. An example could be having all local warehouses in a different country than the central warehouse.

For each network structure the sequence is based on these costs. However, in the case of partial lateral transshipments only one of the other local warehouses with the lowest positive costs is considered.

We considered a large variety of different parameter settings to apply the enumeration procedure. For each of the scenarios we take the replenishment lead time from the supplier to the central warehouse t_0 equal to 12 weeks. The transport time from the central warehouse to the local warehouses (except the emergency warehouse for which this is 0) t_j is equal to 1 week. We vary the demand per week μ_j and holding costs rate per week h as well.

In Table 8 we present the different scenarios, where $\mathbf{S}^*$, and $C(\mathbf{S}^*)$ represent the optimal base stock levels and corresponding costs when we have an emergency warehouse, respectively. $\mathbf{S}^{**}$, and $C(\mathbf{S}^{**})$ represent the optimal base stock levels and corresponding costs when we do not have an emergency warehouse.

Based on these results we can see that the possibility of using the emergency warehouse, and thus keep stock separate for this, does not always lead to a cost reduction. Especially when a lateral transshipment is more interesting cost wise, and the spare parts are not very expensive this rarely will lead to using the emergency warehouse as it is preferable to have more stock at the local warehouses. However, if the emergency warehouse is less expensive than lateral transshipments, or when lateral transshipments are not possible, there is a larger benefit. We even see scenarios for which cost differences can be over 30%. Because the base stock levels have to be integer, it can occur that a small change in the (cost) parameters may lead to a very different outcome. It might thus be that a small change makes the emergency warehouse less beneficial than before or the other way around. These differences are even bigger when there are less options in terms of lateral transshipments. Moreover, when lateral transshipments are cheaper the emergency warehouse becomes less beneficial. The emergency warehouse is especially interesting because of its fast shipments, and thus lower costs for the emergency shipment compared to a shipment from the central warehouse, as well as the ability to deliver the part to every other local warehouse, thus having a pooling effect. If it is then possible to obtain a part by a

Table 8
Differences in costs between having an emergency warehouse or not.

				With emergency warehouse		Without emergency warehouse		
Scenario	μ_j	h	Sequence	$\mathbf{S}^*$	$C(\mathbf{S}^*)$	$\mathbf{S}^{**}$	$C(\mathbf{S}^{**})$	ΔC (%)
1	0,02	150	No-No	(1,0,0,0,0,2)	693,7	(1,1,1,1,1)	886,9	21,78
2	0,02	500	No-No	(1,0,0,0,0,1)	1505,2	(2,0,0,0,0)	1751,2	14,05
3	0,02	1200	No-No	(0,0,0,0,0,0)	2160	(0,0,0,0,0)	2160	0
4	0,1	150	No-No	(5,1,1,1,1,2)	1874,7	(7,1,1,1,1)	2113,9	11,32
5	0,1	500	No-No	(3,0,0,0,0,3)	4780	(5,1,1,1,1)	5625,6	15,03
6	0,1	1200	No-No	(1,0,0,0,0,2)	7910,4	(3,0,0,0,0)	10185	22,33
7	0,2	150	No-No	(8,2,2,2,2,2)	3013,7	(12,2,2,2,2)	3278,6	8,08
8	0,2	500	No-No	(7,0,0,0,0,4)	8359,8	(11,1,1,1,1)	9535,8	12,33
9	0,2	1200	No-No	(3,0,0,0,0,3)	14544	(6,1,1,1,1)	18442	21,14
10	0,02	150	No-Part	(1,0,0,0,0,2)	693,7	(0,1,1,1,1)	773,41	10,31
11	0,02	500	No-Part	(1,0,0,0,0,1)	1505,2	(0,0,0,0,0)	2160	30,31
12	0,02	1200	No-Part	(0,0,0,0,0,0)	2160	(0,0,0,0,0)	2160	0
13	0,1	150	No-Part	(5,1,1,1,1,1)	1771,2	(6,1,1,1,1)	1845,2	4,01
14	0,1	500	No-Part	(3,0,0,0,0,3)	4780	(4,1,1,1,1)	5006,6	4,53
15	0,1	1200	No-Part	(1,0,0,0,0,2)	7910,4	(1,0,1,0,1)	9340,4	15,31
16	0,2	150	No-Part	(8,2,2,2,2,1)	2885,7	(10,2,2,2,2)	2927,4	1,42
17	0,2	500	No-Part	(7,1,1,1,1,2)	8101,4	(7,2,2,2,2)	8608	5,89
18	0,2	1200	No-Part	(3,0,0,0,0,3)	14544	(5,1,1,1,1)	16919	14,04
19	0,02	150	No-Full	(1,0,1,1,0,0)	645,3	(1,0,1,1,0)	645,3	0
20	0,02	500	No-Full	(1,0,0,1,0,0)	1483,2	(1,0,0,1,0)	1483,2	0
21	0,02	1200	No-Full	(0,0,0,0,0,0)	2160	(0,0,0,0,0)	2160	0
22	0,1	150	No-Full	(4,1,1,1,1,1)	1671,8	(5,1,1,1,1)	1707,4	2,09
23	0,1	500	No-Full	(2,1,1,1,1,1)	4405,5	(3,1,1,1,1)	4446	0,91
24	0,1	1200	No-Full	(1,0,0,0,0,2)	7910,4	(1,0,1,1,1)	8321,9	4,94
25	0,2	150	No-Full	(7,2,2,2,2,1)	2795,1	(8,2,2,2,2)	2807,3	0,43
26	0,2	500	No-Full	(6,1,1,1,1,2)	7574,8	(7,1,2,1,2)	7862,8	3,66
27	0,2	1200	No-Full	(3,0,0,0,0,3)	14544	(4,1,1,1,1)	15452	5,88
28	0,02	150	Part-No	(0,1,1,1,1,0)	784,5	(0,1,1,1,1)	784,5	0
29	0,02	500	Part-No	(0,0,1,0,1,0)	1838	(0,0,1,0,1)	1838	0
30	0,02	1200	Part-No	(0,0,0,0,0,0)	2160	(0,0,0,0,0)	2160	0
31	0,1	150	Part-No	(5,1,2,1,2,0)	1916,3	(5,1,2,1,2)	1916,3	0
32	0,1	500	Part-No	(4,1,1,1,1,0)	5159,6	(4,1,1,1,1)	5159,6	0
33	0,1	1200	Part-No	(1,0,1,0,1,0)	9505,3	(1,0,1,0,1)	9505,3	0
34	0,2	150	Part-No	(10,2,2,2,2,0)	2982,2	(10,2,2,2,2)	2982,2	0
35	0,2	500	Part-No	(7,2,2,2,2,0)	8741,9	(7,2,2,2,2)	8741,9	0
36	0,2	1200	Part-No	(4,1,1,1,1,0)	17283	(4,1,1,1,1)	17283	0
37	0,02	150	Part-Full	(0,1,1,1,1,0)	757,3	(0,1,1,1,1)	757,3	0
38	0,02	500	Part-Full	(0,0,1,0,1,0)	1829,7	(0,0,1,0,1)	1829,7	0
39	0,02	1200	Part-Full	(0,0,0,0,0,0)	2160	(0,0,0,0,0)	2160	0
40	0,1	150	Part-Full	(3,2,2,2,2,0)	1909,5	(3,2,2,2,2)	1909,5	0
41	0,1	500	Part-Full	(4,1,1,1,1,0)	5248	(4,1,1,1,1)	5248	0
42	0,1	1200	Part-Full	(1,0,0,0,1,1)	9579,1	(1,0,1,0,1)	9794,1	2,2
43	0,2	150	Part-Full	(10,2,2,2,2,0)	2983,2	(10,2,2,2,2)	2983,2	0
44	0,2	500	Part-Full	(7,2,2,2,2,0)	8752,1	(7,2,2,2,2)	8752,1	0
45	0,2	1200	Part-Full	(5,1,1,1,1,1)	17925	(5,1,2,1,2)	18302	2,06

lateral transshipment from any other local warehouse, there is no longer a big benefit in terms of pooling for the emergency warehouse. However, if parts are very expensive and it is thus too expensive to stock parts locally at multiple locations, the emergency warehouse can still pay off due its fast shipments. Overall it can be concluded that having an emergency warehouse can definitely pay off, although one should consider the structure of the network before deciding on whether it will pay off or not to have this emergency warehouse. If it is faster to obtain parts from other local warehouses compared to the emergency warehouse, it becomes less interesting to have an emergency warehouse, although the emergency warehouse is often organized as such to allow for fast emergency shipments, whereas for local warehouses this is not generally the case. Whereas having local warehouses spread over the continent or world and a centrally located emergency warehouse most likely will be beneficial, unless parts are cheap enough to have abundant amounts of stock at each local warehouse to compensate the downtime costs. As a company generally has to deal with a large variety of different parts with different prices and failure rates, the emergency warehouse could then only be used for those parts for which using an emergency warehouse gives the most benefit.

6. Conclusions

In this paper we consider a two-echelon spare parts network with lateral and emergency shipments, and introduce the use of an emergency warehouse, a warehouse that is able to ship parts fast to any local warehouses in the network in the case of a stockout. We allow for a very general structure at which the lateral and emergency shipments are handled, which enables us to model network structures considered in the literature before as well. We first derive an approximate evaluation procedure that allows us to evaluate the overall performance. By means of simulation we show that the evaluation procedure is accurate. Based on our accurate evaluation procedure we derive a lower bound on the optimal costs, as well as a smart enumeration procedure to find a close-to-optimal solution. We then compare the case where we have the emergency warehouse available to us to the case where we do not have this option available. By comparing the costs of the optimal solutions, we find that costs difference of over 30% are possible. By considering for which instances the cost differences are the largest, companies such as ASML can decide for which parts using such an emergency warehouse is interesting.

For future research it would be interesting to see what happens if we would extend the problem to a multi-item problem where the company has to meet a certain aggregate service level. Although the evaluation would not change, as the parts are still independent, the optimization becomes more complex due to the aggregation of overall performance. A heuristic could be applied, although an incremental policy would most likely not lead to very good results as we have observed that the structure of the solutions change over different ranges of service levels.

Acknowledgments

The authors gratefully acknowledge the support of the Netherlands Organisation for Scientific Research grant project number is 407-12-001.

References

Alfredsson, P., & Verrijdt, J. (1999). Modeling emergency supply flexibility in a two-echelon inventory system. *Management Science, 45*(10), 1416–1431.
Andersson, J., & Melchiors, P. (2001). A two-echelon inventory model with lost sales. *International Journal of Production Economics, 69*(3), 307–315.
Axsäter, S. (1990). Modelling emergency lateral transshipments in inventory systems. *Management Science, 36*, 1329–1338.
Axsäter, S., Howard, C., & Marklund, J. (2013). A distribution inventory model with transshipments from a support warehouse. *IIE Transactions, 45:3*, 309–322.
Axsäter, S., Kleijn, M., & De Kok, T. (2004). Stock rationing in a continuous review two-echelon inventory model. *Annals of Operations Research, 126*, 177–194.
Grahovac, J., & Chakravarty, A. (2001). Sharing and lateral transshipments of inventory in a supply chain with expensive low-demand items. *Management Science, 47*, 579–594.
Graves, S. (1985). A multi-echelon inventory model for a repairables item with one–for-one replenishment. *Management Science, 31*(10), 1247–1256.
Howard, C., Marklund, J., Tan, T., & Reijnen, I. (2015). Inventory control in a spare parts distribution system with emergency stocks and pipeline information. *Manufacturing and Service Operations Management, 17*(2), 142–156.
Karush, W. (1957). A queueing model for an inventory problem. *Operations Research, 5*, 693–703.
Kranenburg, A., & Van Houtum, G. J. (2009). A new partial pooling structure for spare parts networks. *European Journal of Operational Research, 199*(3), 908–921.
Kutanoglu, E. (2008). Insights into inventory sharing in service parts logistics systems with time-based service levels. *Computers and Industrial Engineering, 54*, 341–358.
Kutanoglu, E., & 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*, 728–742.
Muckstadt, J. (2005). *Analysis and algorithms for service part supply chains*. Berlin: Springer.
Muckstadt, J., & Thomas, L. (1980). Are multi-echelon inventory methods worth implementing in systems with low-demand-rate items? *Management Science, 26*, 483–494.
Öner, K., Kiesmüller, G., & Van Houtum, G. (2007). Life cycle costs measurement of complex systems manufactured by an engineer-to-order company. In R. G. Qui, D. W. Russell, W. G. Sullivan, & M. Ahmad (Eds.), *Proceedings of the seventeenth international conference on flexible automation and intelligent manufacturing, FAIM* (pp. 589–596). Philadelphia.
Özkan, E., Van Houtum, G. J., & Serin, Y. (2015). A new approximate evaluation method for two-echelon inventory systems with emergency shipments. *Annals of Operations Research, 224*, 147–169.
Paterson, C., Kiesmüller, G., Teunter, R., & Glazebrook, K. (2011). Inventory models with lateral transshipments: A review. *European Journal of Operational Research, 210*, 125–136.
Reijnen, I., Tan, T., & Van Houtum, G. J. (2009). Inventory planning for spare parts networks with delivery time requirements. Working paper. WP-280, Eindhoven University of Technology, the Netherlands.
Rustenburg, W., van Houtum, G. J., & Zijm, W. (2004). Exact and approximate analysis of multi-echelon, multi-indenture spare parts systems with commonality. In J. Shantikhumar, D. Yao, & W. Zijm (Eds.), *Stochastic modeling and optimization of manufacturing systems and supply chains, Ch. 7* (pp. 143–176). Boston: Kluwer Academic.
Sherbrooke, C. (1968). Metric: A multi-echelon technique for recoverable item control. *Operations Research, 16*(1), 122–141.
Sherbrooke, C. (2004). *Optimal inventory modeling of systems: Multi-echelon techniques*. Wiley.
Van Houtum, G. J., & Kranenburg, A. A. (2015). *Spare parts inventory control under system availability constraints*. New York: Springer.
van Wijk, A., Adan, I., & Van Houtum, G. J. (2012). Approximate evaluation of multi-location inventory models with lateral transshipments and hold back levels. *European Journal of Operational Research, 218*(3), 624–635.
van Wijk, A., Adan, I., & Van Houtum, G. J. (2013). Optimal allocation policy for a multi-location inventory system with a quick response warehouse. *Operations Research Letters, 41*(3), 305–310.
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, 56*, 1419–1430.
Wong, H., Kranenburg, B., Van Houtum, G. J., & Cattrysse, D. (2007). Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse. *OR Spectrum, 29*, 699–722.