Int. J. Production Economics ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Int. J. Production Economics

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

Allocating service parts in two-echelon networks at a utility company

D. van den Berg [a,*], M.C. van der Heijden [b], P.C. Schuur [b]

[a] Alliander, Utrechtseweg 68, 6812 AH Arnhem, The Netherlands
[b] University of Twente, School of Management and Governance, P.O. Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 30 September 2014
Accepted 26 August 2015

Keywords:
Service parts
Multi-echelon
Lost sales
Inventory allocation
Batching

ABSTRACT

We study a multi-item, two-echelon, continuous-review inventory problem at a Dutch utility company. We develop a model for the optimal allocation of service parts in a two-echelon network under an aggregate waiting time constraint. Specific model aspects are emergency shipments in case of stockout, and batching for regular replenishment orders at the central warehouse. We use column generation to solve this problem with various building blocks for single-item models as columns. Further, we derive simple classification rules from the solution of our multi-item, two-echelon service part optimization problem using statistical techniques. Application of our models at Liander yields a cost reduction of 15% and a decrease in the impact of waiting time for service parts by 52%.

1. Introduction

In the utility sector, uptime of the network is important to provide the consumers gas and electricity according to their needs. This requires a quick response to failures, and so the resources needed should be readily available. In this study, we focus on service part provisioning for corrective maintenance of the energy network of a Dutch utility company, Liander. Liander distributes electricity and gas to a third of the Netherlands and has 3 million electricity and 2.1 million gas connections. Liander owns the grid from the connection at people's home to upstream in the grid (electricity until 50 kV, gas until 8 bar upstream). Liander controls 87,483 km of electricity network and 42,460 km of gas network. In this research, we focus on the electricity network.

The impact of defective service parts varies greatly: a meter may cause an outage for one connection while a big transformer may cause an outage for 1000 connections. To minimize the outage, Liander stocks about 5400 service part types, such as meters, cables, transformers, switchboards and sockets. The part value varies from €0.01 to €35,000.-. The total inventory value for service parts was approximately €5 million at the start of our project. A key performance indicator for network availability is the *relative minutes of downtime* (*RMD*). This measure expresses the average number of minutes in a year that each electricity connection was down in Liander's service area. Given F failures in a year with failure f lasting for t_f minutes and affecting c_f out of the C connections, we have that $RMD=\sum_{f=1}^{F} c_f t_f / C$. Liander achieves on average an *RMD* between 20 and 30 min.

The availability of service parts is crucial to attain a low *RMD*. Liander estimates that currently the service part deficiency contributes approximately 10% to the total *RMD*. Next, service part deficiency also impacts planned maintenance and projects, as unforeseen demand for service parts causes idle time for mechanics. Both power outages and unforeseen demand arising from planned maintenance and projects are covered in this research as *urgent orders*. We include both types of impact in a key performance indicator for our study, being the Minutes impact of Service Parts Deficiency (*MSPD*), see Section 3.1 for details. The supply chain of Liander consists of suppliers, a central warehouse and several regional manned and unmanned warehouses, see Fig. 1.1.

At the start of the project, inventory control was decentralized. Every warehouse had high safety stocks for fast movers while certain expensive slow movers were not stocked anywhere. Some stock was allocated at the supplier as consignment stock. Liander initiated a project, centralizing inventory control for urgent order fulfillment with the aim to deliver higher service levels for less inventory costs. For each item, we should determine at which locations and in what quantities it should be stocked, based on a trade-off between *MSPD* on the one hand, and inventory and transportation costs on the other hand. The supply chain is simplified by considering four network options per item:

- *Network 1-consignment stock at the supplier only.*
- *Network 2-stocking at the central warehouse only*.

* Corresponding author.
E-mail addresses: diederickvdberg@gmail.com (D. van den Berg), m.c.vanderheijden@utwente.nl (M.C. van der Heijden), p.c.schuur@utwente.nl (P.C. Schuur).

http://dx.doi.org/10.1016/j.ijpe.2015.08.025
0925-5273/

- *Network 3-stocking at the central and all regional manned warehouses*. The regions assigned to unmanned warehouses are supplied by the nearest manned warehouse.
- *Network 4-stocking at the central and all regional warehouses.*

If a location is out of stock when demand occurs, an emergency order is placed at the next location upstream in the supply chain (central warehouse, supplier), see Table 1.1. If a service part is not available in the network, we assume that an *infinite capacity supplier* can always supply the part at high costs and long shipment times. In practice, this is the supplier without consignment stock. Note that lateral shipments (i.e., mechanics being supplied from another regional warehouse than their own) are considered to be impractical by Liander and therefore excluded from the model, even though it could be fitted in the optimization approach.

Expensive slow movers should typically be stored upstream in the supply chain (Network 1 or 2), whereas Network 3 and 4 are more suitable for cheap fast movers. The distinction between Network 3 and 4 is that the manned warehouses in Network 3 profit from risk pooling, but also cause longer average shipment times to customers as they serve larger geographical areas. There are 27 regional warehouses, 9 manned and 18 unmanned. Liander requires uniformity of assortment, resulting in a minimal stock level of one for each warehouse being used. Therefore, Network 3 is not a special case of Network 4, and Network 2 is not a special case of Network 3. Network 1 has maximum risk pooling effect, as the inventories may be shared with other customers of the supplier, and Liander only contracts suppliers who are not too much dependent on Liander. Also, the supplier can stock components from which it can quickly assemble different service parts. Compared to Network 2, the inventory reduction comes at the price of longer shipment times and higher shipment costs. For Network 1, we assume that we contract a certain fill rate with the supplier, valid for all parts. Fill rate differentiation is less useful since Network 1 will mainly be used for expensive slow movers.

For each service part, we should select a *delivery policy*, consisting of a network and the inventory levels per site. We aim to find a set of delivery policies that minimizes the total costs of inventory holding and emergency shipments such that a target value for *MSPD* is not exceeded.

The remainder of the paper is structured as follows. In Section 2, we describe the relevant literature and state our contribution. Section 3 describes our model and the solution approach. Section 4 applies our model using Liander's data. As Liander prefers to have a simple and intuitively logical framework for the choice of the delivery policies, we develop such a framework in Section 5. There, we also examine the cost penalty of replacing an advanced optimization routine by an item approach. Finally, we draw our conclusions in Section 6.

2. Literature

In the past decades, a huge amount of literature has been published on spare part inventory models (Basten and van Houtum, 2014). The seminal paper in this area is Sherbrooke (1968) on multi-item, multi-echelon spare part stocking. A lot of related work is covered in Sherbrooke (2004) and Muckstadt (2005). Many models assume backordering if a spare part is not on stock. When however the consequences of system downtime are huge in terms of costs or reduced quality of service, alternative sourcing options are typically considered, such as emergency shipments from locations upstream in the supply chain. Since the corresponding demand does not need to be satisfied anymore from the original stockpoint, this is typically modeled as a lost sales inventory model.

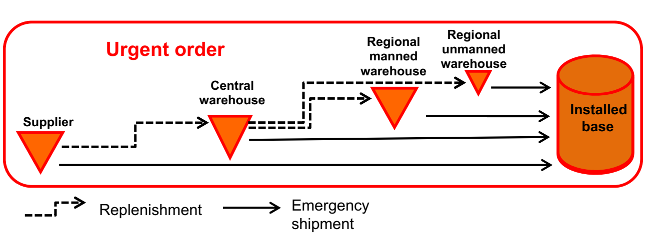

Fig. 1.1. Graphical representation of the supply chain of Liander.

Table 1.1
Per network the sequence in which the stockpoints are requested for delivery to the site of need.

	Infinite capacity supplier	Consignment stock	Central warehouse	Regional warehouse	
				Manned	Unmanned
Network 1	2	1	–	–	–
Network 2	2	–	1	–	–
Network 3	3	–	2	1	–
Network 4	3	–	2	1	1

An overview of inventory models with lost sales is given in Bijvank and Vis (2011). Their focus is primarily on single site inventory systems which are useful for Network 1 and 2 only (cf. Table 1.1). Muckstadt and Thomas (1980) consider a two-echelon model for a single item in which demand that cannot be satisfied from stock at a local warehouse is transmitted to the central depot. If stock is depleted there as well, the demand is satisfied from an external source with unlimited capacity, and so the demand is lost to the system. They approximate the demand process at the central depot by a Poisson distribution. Their method has been improved by Özkan et al. (in press). Alfredsson and Verrijdt (1999) combine lateral transhipments between local warehouses and emergency shipments from upstream in the supply chain in their model, where lateral transhipments have priority. Andersson and Melchiors (2001) consider a similar model to Muckstadt and Thomas (1980), where demand at a local warehouse is lost if the warehouse is out of stock, even when there is still stock at the central depot. Alvarez and van der Heijden (2014) consider a variant of the model by Muckstadt and Thomas (1999) with short regular replenishment lead time between central depot and local warehouse, so that emergency shipments are not initiated if an item in the pipeline to the local warehouse is still available. For our application, the model by Özkan et al. (in press) is most suitable as a single-item building block. As their assumption of one-for-one replenishment is not applicable for the fast movers in our application, we will derive a variant with lot sizing at the central depot.

In case of backordering, the "biggest-bang-for-the-buck" heuristic by Sherbrooke (1968) is typically used. For lost sales models, a suitable approach is column generation, see e.g. Wong et al. (2007), Kranenburg and van Houtum (2007, 2008) and Alvarez et al. (2013, in press). The power of this approach is its flexibility: we can embed a large variety of single inventory models in the multi-item optimization, as long as the performance evaluation per item policy is fast and accurate. The drawback is that it may be computationally burdensome if the number of items is high. We will solve this by splitting the parts in two sets: a set of expensive and critical slow movers for which accurate stock levels are important to find, and a set of cheap fast movers for which we should typically not run out of stock, since downtime costs are far higher than inventory holding costs.

Only few papers have been published on case studies in multi-item, multi-echelon spare part optimization. Cohen et al. (1990) describe a spare part optimization tool for IBM, focused on fast

Please cite this article as: van den Berg, D., et al., Allocating service parts in two-echelon networks at a utility company. International Journal of Production Economics (2015), http://dx.doi.org/10.1016/j.ijpe.2015.08.025

moving items controlled by (s, S) inventory policies. The target part availability is not a model outcome as we aim for, but should be specified as input. Greedy heuristics are developed to solve large scale problems with up to 200,000 parts. Korevaar et al. (2007) describe an application at a German automobile manufacturer, where the complexity is limited to single-echelon systems. Şen et al. (2010) describe a mathematical program for the design of the service and parts network of Applied Materials, where inventories are roughly modeled, as the related cost is just one of the factors influencing network design. It is assumed that all parts at a certain location have the same service level, whereas we aim to differentiate service levels.

Regarding spare part classification, Rossetti and Achlerkar (2011) describe a methodology which groups items, and next sets inventory policies in multi-item, single location inventory systems. Hopp et al. (1999) present an "easily implementable" approach in a two-echelon system which is able to handle changes in assortment and only requires resolving the entire problem in case of major changes. However, the nature of the problem is different due to the lack of emergency shipments, such that we need to develop a new approach that covers both network choice and inventory levels.

In summary, we aim to contribute to the literature as follows:

I. We solve a multi-item, two-echelon service part optimization problem by combining several building blocks from literature and show its added value in practice.
II. We extend the model of Özkan et al. (in press) with replenishment lot size larger than one at the central depot.
III. We build a framework for simplifying service part inventory decisions for the case company based on a statistical analysis of the results from the multi-item, two-echelon optimization model and show the cost of simplification in a numerical experiment.

3. Model description

In Section 3.1, we outline our model. Section 3.2 describes the notation and our optimization model. Section 3.3 specifies the performance evaluation of a delivery policy per network type. Section 3.4 describes the column generation approach for critical, high value items. We develop a separate approach-based on the results from column generation-to optimize low value items in Section 3.5.

3.1. Outline

First, let us define our key performance indicator, the Minutes impact of Service Parts Deficiency (*MSPD*). Let I denote the set of service parts, D_i the average demand for part i per year, mi_i the minutes impact due to service part i, and EW_i the average waiting time for part i. Then we have:

$$MSPD = \sum_{i \in I} D_i mi_i EW_i \tag{1}$$

The minutes impact mi_i of item i consists for a fraction α_i of demand resulting from a power outage, and a fraction $(1-\alpha_i)$ of idle time of a mechanic. We define c_i as the number of connections affected by a power outage due to deficiency of service part i, and C as the total number of connections. Also, we rate the idle time of a service engineer as being equal to itm connections affected. Then,

$$mi_i = \frac{\{\alpha_i c_i + (1-\alpha_i) itm\}}{C.} \tag{2}$$

We can influence *MSPD* via the average waiting time for each part, which is determined by the network structure and the stock levels at its corresponding warehouses. We aim to minimize the cost of inventories and emergency shipments, such that a target value for *MSPD* is not exceeded. When in Network 3 and 4 demand at a regional warehouse cannot be satisfied from stock on shelf, an emergency shipment is issued from the central warehouse (CW) if it still has the item on stock, irrespective of items in the regular replenishment pipeline between central and regional warehouse. If the CW is out of stock as well, an emergency shipment from the infinite capacity supplier is issued, irrespective of items in the regular replenishment pipeline between supplier and CW. This is the current modus operandi at Liander, since the regular replenishment lead times, minimally three days, are unacceptably long in case of urgent demand. Obviously, the emergency shipment time is less than the regular replenishment lead time from the same stockpoint.

In Network 3 and 4 we model the lead time from supplier to the central warehouse by an exponential distribution. This assumption facilitates Markov chain analysis, and Alfredsson and Verrijdt (1999) have shown that the performance of these networks is rather insensitive to the lead time distribution if the central warehouse uses one-for-one replenishment. In Section 4.1 we separately assess the impact of the lead time distribution if the central warehouse uses a replenishment quantity larger than one. We use the Erlang loss formula for performance evaluation of Network 1 and 2. As these networks will typically be selected for expensive slow movers, a replenishment lot size of one makes sense. Further, we use the following model assumptions:

1. The demand is Poisson distributed and independent across items and regions.
2. All regional warehouses use an $(S\text{-}1,S)$-policy, implying continuous review and a replenishment order size of one; the same policy applies for consignment stock at the supplier in Network 1.
3. The central warehouse uses an (s,Q)-policy, implying continuous review with a reorder point of s and a fixed replenishment order size of $Q \geq 1$. The order size Q is predetermined using the well-known economic order quantity formula.
4. In Network 1, all service parts have the same fill rate which is given as model input.

3.2. Notation

We consider items $i \in I$ in delivery networks $g \in G=\{1,2,3,4\}$. We index the stockpoints (referred to as warehouses) by $n \in N \equiv N \cup \{\infty\}$, where index ∞ refers to the infinite capacity supplier, 0 refers to consignment stock, and index 1 refers to the central warehouse. Indices 2,.., M and $M+1$,.., $M+U$ refer to the manned and unmanned regional warehouses, respectively. L_g denotes the set of regional warehouses in N for network g, where $L_1 = \varnothing$, $L_2 = L_4 = \{2,\ldots, M+U\}$ and $L_3 = \{2,\ldots,M\}$. Although we use no regional warehouses in Network 2, we add them to L_2 such that we can develop uniform expressions later on in this section. Let R denote the set of demand regions, $r \in R$. We start indexing r at 2 such that it runs parallel with the regional warehouses.

Input parameters

TE_{nr} Average (emergency) shipment time from warehouse n to region r. In Network 1 we use TE_{0r} and $TE_{\infty r}$, in Network 2 TE_{1r} and $TE_{\infty r}$ and in Network 3 and 4 TE_{nr}, TE_{1r} and $TE_{\infty r}$, $n \in L_g$, $g \in \{3, 4\}$.

Please cite this article as: van den Berg, D., et al., Allocating service parts in two-echelon networks at a utility company. International Journal of Production Economics (2015), http://dx.doi.org/10.1016/j.ijpe.2015.08.025

LT_{in} Replenishment lead time of warehouse $n\epsilon\{0,\ldots,M+U\}$ for item i.
CE_{nr} Additional costs of an emergency shipment to region r from warehouse $n \in \{\infty, 0, 1\}$ compared to a delivery from a regional warehouse. In general, $CE_{\infty r} > CE_{0r}$.
d_{ir} Demand rate from region r for item i per period (e.g. year), $D_i = \sum_{r\in R} d_{ir}$.
LK_{nrg} Binary parameter indicating the link between region and regional warehouse, i.e., $LK_{nrg}=1$ if region r is linked to warehouse n in network g and $LK_{nrg}=0$ otherwise, $g \in G\{1\}$.
h_i Holding costs for item i.
Q_{in} Fixed lot size for item i at warehouse n.
α_i Fraction of demand for item i arising from power outages.
c_{ir} Average number of connections affected by a power outage due to item i in region r.
C Total number of connections.
itm Idle time of mechanics expressed as an equivalent number of affected connections.
mi_{ir} Minutes of impact when item i in region r is not available.
mx Maximum impact allowed due to service parts deficiency ($MSPD$) over all items.
FS_i Fraction of item i demand at the supplier with consignment stock originating from Liander.

Decision variables

g_i Network for item i.
$\boldsymbol{s}_{ig_i}$ Vector of reorder points for item i in network g_i, with s_{in}=reorder point at warehouse n, and $S_{in} = s_{in} + Q_{in}$. When warehouse n is not part of network g_i, it holds that $s_{in}=Q_{in}=0$.
$\boldsymbol{p}_i$ Delivery policy $(g_i, \boldsymbol{s}_{ig_i})$, a combination of the decision variables above.

Auxiliary variables

$\beta_{in}(\boldsymbol{p}_i)$ Fill rate of warehouse n ϵ $\{0,\ldots,M+U\}$ using delivery policy $\boldsymbol{p}_i$ for item i.
$\theta_{in}(\boldsymbol{p}_i)$ Fraction of demand satisfied by an emergency shipment from the central warehouse at regional warehouse $n \epsilon L_{g_i}$ using delivery policy $\boldsymbol{p}_i$, $g_i \in G\{1\}$ for item i.
$\gamma_{in}(\boldsymbol{p}_i)$ Fraction of demand satisfied by an emergency shipment from the infinite capacity supplier at regional warehouse $n \epsilon L_{g_i}$ using delivery policy $\boldsymbol{p}_i$, $g_i \in G\{1\}$ for item i·

For Network 2-4, it holds that Obviously, $\beta_{in}(2, \boldsymbol{s}_{i2})=0$ $\forall\ n \in L_2,\ i \in I$.

Performance indicators (output)

$TC_i(\boldsymbol{p}_i)$ Total cost of item i using delivery policy $\boldsymbol{p}_i$ (holding and emergency shipment costs).
$EW_{ir}(\boldsymbol{p}_i)$ Expected waiting time of item i at region r using delivery policy $\boldsymbol{p}_i$.

The optimization problem can now be expressed as **Problem I**:

$$\min_{p_i} \sum_{i=1}^{|I|} TC_i(\boldsymbol{p}_i)$$
$$s.\,t.\ \sum_{i=1}^{|I|}\sum_{r\in R} d_{ir} EW_{ir}(\boldsymbol{p}_i) mi_{ir} \le mx \tag{3}$$

In restriction (3), we use the definition of $MSPD$ as in (1) and (2). The expected costs of item i are:

$$TC_i(\boldsymbol{p}_i) = \begin{cases} S_{i0}FS_i h_i + \sum_{r\in R} d_{ir}\left\{\beta_{i0}(\boldsymbol{p}_i) CE_{0r} + \left(1-\beta_{i0}(\boldsymbol{p}_i)\right) CE_{\infty r}\right\} & g_i = 1 \\ h_i\left(s_{i1} + \frac{Q_{i1}+1}{2}\right) + \sum_{n\in L_{g_i}}\left\{S_{in} h_i + \sum_{r\in R} d_{ir} LK_{nrg_i}\left(\theta_{in}(\boldsymbol{p}_i) CE_{1r} + \gamma_{in}(\boldsymbol{p}_i) CE_{\infty r}\right)\right\} & g_i \in \{2,3,4\} \end{cases} \tag{4}$$

The inventory costs are determined by the stock in the warehouses of a network. Liander becomes owner of the stock at the moment of ordering at the supplier. In Network 1, Liander pays only for the consignment stock proportional to the size of its demand at the supplier. In the other networks, we take into account the replenishment size at the central warehouse and link the demand of a region to an (un)manned warehouse by LK_{nrg}. Next, we determine the expected costs of the emergency shipments based on the values θ_{in}, γ_{in}. The same logic is applied to the expected waiting time:

$$EW_{ir}(\boldsymbol{p}_i) = \begin{cases} \beta_{i0}(\boldsymbol{p}_i) TE_{0r} + \left(1-\beta_{i0}(\boldsymbol{p}_i)\right) TE_{\infty r} & g_i = 1 \\ \sum_{n\in L_{g_i}} LK_{nrg_i}\left(\beta_{in}(\boldsymbol{p}_i) TE_{nr} + \theta_{in}(\boldsymbol{p}_i) TE_{1r} + \gamma_{in}(\boldsymbol{p}_i) TE_{\infty r}\right) & g_i \in \{2,3,4\} \end{cases} \tag{5}$$

The expected waiting time for an item is a weighted average of the (emergency) shipment times from warehouse to region with the sourcing fractions as weights. As stated before, we differ between shipment time from the infinite capacity supplier and the supplier with consignment stock.

Now we reformulate Problem I to an equivalent Problem II which we can solve by column generation. This approach enables decomposition of the problem into single-item problems and allows us to include non-linear aspects in a multi-item problem. We proceed from subsets of policies P_i that can be used for item i. The decision becomes then to select the right delivery policy p_i from P_i such that we do not exceed the target value mx, and minimize the total costs. The binary decision variable X_{ip_i} is 1 if policy $p_i \epsilon P_i$ is selected for item i and 0 otherwise. **Problem II** becomes:

$$\min_{X_{i\boldsymbol{p}_i}} \sum_{i=1}^{|I|}\sum_{\boldsymbol{p}_i\in \boldsymbol{P}_i} X_{i\boldsymbol{p}_i} TC_i(\boldsymbol{p}_i)$$
$$s.\,t.\ \sum_{i=1}^{|I|}\sum_{r\in R}\sum_{\boldsymbol{p}_i\in \boldsymbol{P}_i} X_{i\boldsymbol{p}_i} d_{ir} EW_{ir}(\boldsymbol{p}_i)^{*} mi_{ir} \le mx \tag{6}$$

$$\sum_{\boldsymbol{p}_i\in \boldsymbol{P}_i} X_{i\boldsymbol{p}_i} = 1\ \ \forall\, i$$
$$X_{i\boldsymbol{p}_i} \in \{0,1\}\ \ \forall\, i,\ \boldsymbol{p}_i \in \boldsymbol{P}_i \tag{7}$$

In order to find a near optimal set of item policies for each item, we solve the LP-relaxation of Problem II, and derive from its

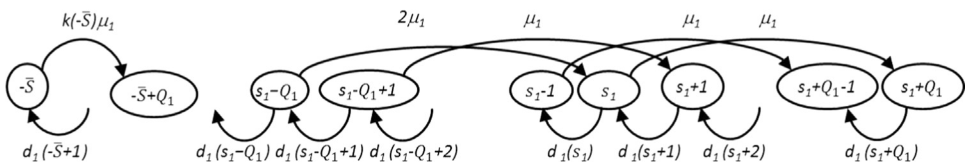

Fig. 3.1. Graphical illustration of the Markov chain for the inventory level at the central warehouse.

solution which alternative policies not included in P_i should be added in the next iteration, until no promising policy can be found anymore for any item. We initially create subsets P_i such that a feasible solution exists, see Section 3.4.1. From the shadow prices of the two constraints, obtained after solving the LP-relaxation, we determine whether an unconsidered policy has negative reduced costs, i.e., has the potential to improve the objective function. In Section 3.4.2 we furthermore limit the policies for explicit evaluation in a column generation step using the approach of Alvarez et al. (in press). The stopping criterion is when no new policies with negative reduced costs can be found. Section 3.4.3 describes how to obtain an integer solution, but first we show how we evaluate the delivery policies.

3.3. Evaluation of a delivery policy

For ease of explanation, we first discuss Network 1 and 2, and next Network 3 and 4. For notational convenience, we omit suffix i here.

3.3.1. Network 2 and 1

Recall that we assume $Q_1=1$ for Network 2, as stocking service parts at a single central location is likely to be used for low demand, high costs items. The demand rate at the central warehouse is the sum of the regions' arrival rates. The fraction of demand satisfied by the central warehouse, θ_n, with reorder level s_1, is given by the Erlang loss formula with $\rho_1 = DLT_1$:

$$\theta_n = 1 - ERL(S_1, \rho_1) = 1 - \frac{\frac{\rho_1^{S_1}}{S_1!}}{\sum_{j=0}^{S_1} \frac{\rho_1^j}{j!}} \quad \forall\, n \in L_2 \tag{8}$$

with $\gamma_n = 1 - \theta_n$. For Network 1, we assume the supplier to offer a fixed fill rate β_0. We find the base stock level by the Erlang loss formula as stated above, where we increase the demand at the supplier by $\frac{1}{FS}$ (recall that FS denotes the fraction of demand at the supplier that is generated by Liander).

3.3.2. Network 3 and 4

Network 3 and 4 only differ in the number of regional warehouses. The model by Özkan et al. (in press) fits best to the current way of working at Liander (cf. Section 2). The approximation procedure distinguishes three cases from where a demand is fulfilled: (i) regional warehouse, (ii) central warehouse, and (iii) central repair facility. We see the central repair facility as an infinite capacity supplier. The approximation procedure models the inventory position of the central warehouse by a Markov chain. It iteratively determines the fill rate at the regional warehouses given a delay in lead time from the central warehouse and determines the delay at the central warehouse given fill rates for the regional warehouses. After these values converge, the fractions from the central warehouse and supplier are determined. To allow a replenishment order size $Q_1 > 1$, we define:

IL_1 Inventory level at the CW: inventory on hand minus backorders.

$-\bar{S}$ Lower bound on the inventory level of the CW, $\bar{S} = \Sigma_{n\in L_g} S_n$ $g \in \{3, 4\}$.

IP_1 Inventory position at the CW.

The demand is now suffixed by the regional warehouse. The demand at the CW becomes: $d_1(x)$=demand rate at the CW when its inventory level is x:

$$d_1(x) = \begin{cases} \sum_{n\in L_g} d_n & g \in \{3,4\} \text{ if } x > 0 \\ \sum_{n\in L_g} \beta_n d_n & g \in \{3,4\} \text{ if } x \le 0 \end{cases}$$

Let the lead time of the CW be exponentially distributed with rate μ_1. Since $IP_1 = IL_1 +$quantity on order ($=k^*Q_1$ with $k \in N_0$), we know that:

- There is no outstanding order if $IL_1 > s_1$
- There is exactly one outstanding order if $s_1\text{-}Q_1 < IL_1 \le s_1$
- There are k outstanding orders if $s_1\text{-}kQ_1 < IL_1 \le s_1\text{-}(k-1)Q_1$

More specifically, if the inventory level or state is x ($-\bar{S} \le x \le s_1 + Q_1$), the number of outstanding orders equals: $k(x) = \left\lfloor \frac{s_1 + Q_1 - x}{Q_1} \right\rfloor$, where $\lfloor y \rfloor$ denotes the largest integer smaller than or equal to y. We can modify the Markov chain as shown in Fig. 3.1 of Özkan et al. (in press) accordingly. That is, at a certain inventory level x, we know how many orders are outstanding. The arrival rate of these replenishment orders is $k(x)\mu_1$. With this rate, the Markov chain jumps from inventory level $IL_1=x$ to $IL_1=x+Q_1$. The lower bound $-\bar{S}$ on the inventory level still applies.

The balance equations, specified by the relations *rate in=rate out* $\{x, x+1, \ldots, s_1+Q_1\}$, are:

$$\pi_x = \frac{\sum_{y=max\{x-Q_1,-\bar{S}\}}^{x-1} k(y)\mu_1\pi_y}{d_1(x)} \tag{9}$$

We find the state probabilities π_x using the normalization equation $\Sigma_x\, \pi_x=1$, see Appendix A for the calculation procedure. Appendix B describes our adaptation of Özkan's procedure. We know that for inventory systems with one-for-one replenishment and lost sales the system performance is rather insensitive to the lead time distribution. In Section 4.1 we show that this insensitivity is less for lot sizes larger than one, but remains within reasonable limits for practical purposes.

3.4. Column generation approach

3.4.1. Initial set of item policies

To find an initial policies set guaranteeing a feasible solution to the LP-relaxation, we choose for each item: (i) a set of policies profiting from the risk pooling effect (Network 1 or 2), and (ii) a policy that yields minimum waiting time (Network 4 with fill rate close to 1, as the average distance to the customer is smallest). In Network 2, we find an upper bound s_{i1}^{UB} on the reorder point s_{i1} from $\theta_{in}(2,s_{i1}^{UB})\ge 1 - \varepsilon$, where we set $\varepsilon=0.0001$. Subsequently, we add delivery policies for item i with the following reorder points: $s_{i1}\in\{0, \ldots, s_{i1}^{UB}\}$. To ensure that the required service level mx can be reached, we add one Network 4 option per item with base stock level S_{in} satisfying a fill rate $\beta_{in} \ge 1-\varepsilon$. We use zero stock at the central warehouse and calculate the fill rate by the Erlang loss formula (8), where the replenishment time becomes $LT_{in} + LT_{i1}$, cf. Alvarez et al. (in press).

3.4.2. Generation of new columns

By column generation we search for new delivery policies with negative reduced costs in Problem II (Section 3.2), as these policies will improve the LP solution. We do not have to consider additional policies from Network 1 and 2, as these are in the initial policy set. In every iteration, we add the policy with minimal reduced costs per item, insofar negative. We continue until we cannot find new policies anymore. Let us define $A \le 0$ as the shadow price of the *MSPD* constraint (6), and $J_i \ge 0$ as the shadow price of the policy constraint (7), for item *i*. Then new delivery policy p_i have reduced costs:

$$red_i(\boldsymbol{p}_i) = TC_i(\boldsymbol{p}_i) - J_i - \sum_{r \in R} EW_{ir}(\boldsymbol{p}_i) * mi_{ir} * d_{ir} * A \quad (10)$$

An important part of the approach is to limit the number of policies to be considered. To this end, we follow Alvarez et al. (in press). They formulate three observations which help us to create upper and lower bounds on the stock levels of the warehouses. Furthermore, they conclude empirically between what stock levels the optimal stock level of a regional warehouse would be, given a proposed stock level of the central warehouse. With these bounds, a delivery policy with minimal reduced costs is found rather quickly. For an extensive discussion, we refer to Alvarez et al. (in press). For the adaptations we needed to make, we refer to Appendix C.

3.4.3. Obtaining an integer solution

When a solution is found to the LP-relaxation, we exclude all dominated policies from the P_i, in order to speed up the computation time of the ILP problem. Dominated policies have both a higher average waiting time and higher total costs than at least one other policy of the same item, cf. Alvarez et al., in press. We solve the ILP problem by the CPLEX solver, version 12.2. An alternative is to use branch-and-price with branching on a variable $X_{i\boldsymbol{p}_i}$ that has a noninteger value in the optimal solution of the LP-relaxation. In each iteration, we check whether new columns with negative reduced costs can be found, and repeat until the LP-relaxation finds an integer solution for all remaining branches. We chose for the first option because it proved to work well on similar problems (Alvarez et al., in press). Also, we found a very small gap between the solution value of the ILP problem and the LP-relaxation, see Section 4.3.1.

3.5. Low valued items

Most items have a value below €100 (4249 out of the roughly 5400). Optimization of all these items would require a long computation time. Furthermore, it is obvious that Network 4 is the most suitable for cheap items, as it facilitates short waiting times at low costs. A complication is that our assumption of one-for-one replenishment at the regional warehouses is not valid for cheap items. Therefore, we first calculate the reorder quantity: $Q_{in} n \in L_{g_i}$ by the *EOQ* formula. Next, we apply the fill rate formula assuming backordering, as it does not differ a lot from the lost sales case considering that the fill rate is typically very high for fast movers:

$$\frac{1}{Q_{in}} \sum_{y=s_{in}+1}^{s_{in}+Q_{in}} F(y-1) = \beta_{in} \quad (11)$$

where $F(y)$ is the Poisson distribution with as mean the multiplication of demand rate and replenishment time: $\beta_{i1} LT_{in} + (1-\beta_{i1})(LT_{in} + LT_{i1})$, assuming a high fill rate at the central warehouse, β_{i1}, see Table 4.2. This formula is based upon the fact that the inventory position has a discrete uniform distribution on $\{s_{in}+1,\ldots, s_{in}+Q_{in}\}$, see Axsäter (2006). The costs and waiting times are calculated similar to expressions (4) and (5), assuming that the central warehouse reaches a fill rate of 100% for emergency shipments. In a rare event of a stock out at the *CW*, the item can frequently be obtained at a builder's merchant. Next, we apply Expression (10) to find the reorder point that gives minimum reduced costs. We omit shadow price J_i as it doesn't affect the choice of the reorder point. We use the shadow price *A* of the *MSPD* constraint (6) that we found in the last iteration of the column generation procedure, representing the optimal balance in waiting time and costs.

4. Model application at Liander

In Section 4.1, we evaluate our adapted method of Özkan et al. (in press). Next, we describe the data for the Liander case (Section 4.2 for the model). In Section 4.3 we present the results of our case.

4.1. Evaluation adapted method for $Q_1 > 1$

To assess the accuracy of our approximations, we constructed a simulation model in Plant Simulation software. Table 4.1 displays the average and maximum absolute deviation between our adapted approach of Özkan et al. (in press) and the results from simulation with deterministic lead times. The test instances are derived from Özkan et al. (in press), see Appendix D in the online

Table 4.1
The average and maximum of the absolute deviations between the outcomes of our adapted method of Özkan et al. (in press) and the simulation results.

# Reg. warehouses		Average					Maximum				
		2	4	10	20	Avg.	2	4	10	20	Max
β	$Q_1=1$	0.008	0.008	0.006	0.004	0.007	0.031	0.026	0.027	0.020	0.031
	$Q_1=5$	0.015	0.016	0.023	0.020	0.018	0.047	0.055	0.062	0.057	0.062
	$Q_1=10$	0.011	0.014	0.025	0.027	0.019	0.038	0.050	0.067	0.071	0.071
Avg. and max:		0.011	0.013	0.018	0.017	0.015	0.047	0.055	0.067	0.071	0.071
θ	$Q_1=1$	0.025	0.014	0.008	0.005	0.013	0.080	0.045	0.022	0.013	0.080
	$Q_1=5$	0.006	0.006	0.004	0.006	0.005	0.018	0.022	0.019	0.020	0.022
	$Q_1=10$	0.004	0.008	0.010	0.007	0.007	0.010	0.031	0.034	0.028	0.034
Avg. and max:		0.012	0.009	0.007	0.006	0.008	0.080	0.045	0.034	0.028	0.080
γ	$Q_1=1$	0.023	0.013	0.006	0.004	0.011	0.058	0.032	0.017	0.012	0.058
	$Q_1=5$	0.015	0.017	0.022	0.019	0.018	0.059	0.069	0.061	0.053	0.069
	$Q_1=10$	0.012	0.021	0.035	0.034	0.025	0.047	0.059	0.098	0.082	0.098
Avg. and max:		0.017	0.017	0.021	0.019	0.018	0.059	0.069	0.098	0.082	0.098

Please cite this article as: van den Berg, D., et al., Allocating service parts in two-echelon networks at a utility company. International Journal of Production Economics (2015), http://dx.doi.org/10.1016/j.ijpe.2015.08.025

supplement, where we add instances with $Q_1=5$ and $Q_1=10$ and show detailed approximation and simulation results.

Table 4.1 shows that the average deviations for all three fractions are below 0.025, but we do observe some outliers of 0.098 and 0.08. Interestingly, the value 0.08 corresponds to $Q_1=1$ and is therefore not due to our model extension. The large deviations for $Q_1>1$ are all overestimates of γ_n. Actually, for $Q_1>1$ our adapted method almost consequently overestimates γ_n and underestimates β_n, resulting in a conservative performance estimation. For β_n and γ_n the deviations increase with the lot size. The large deviations occur for $LT_1=20$. At Liander, the maximum LT_1 is 16, the next largest LT_1 is 12 and all other values for LT_1 are below 8. For all items at Liander it holds that $Q_1\le 10$. From these observations, we conclude that the accuracy is sufficient for the purpose of this study.

4.2. Establishing model input

Demand forecast – A long history of local energy company mergers, with their own grid composition, led to the present distribution grid of Liander. Therefore, there is a geographical variety in the parts. This variety causes regional warehouses to have their own unique demand characteristics. We estimated demand characteristics based on data on the replenishments of the regional warehouses and the emergency shipments from the central warehouse for a period of respectively three and two years. For many items, we found a zero demand forecast in several regions. Actually, this is due to statistical fluctuations: these demand rates will be low but strictly positive. For these items, we estimated the aggregate demand over all regions, and allocated this demand to the regions based on an estimation of the installed base per item per region. In case of an emergency shipment, some items are ordered in fixed quantities larger than one. We simply adapted the unit size definition accordingly, such that the assumption of unit sized demand remains valid.

Item criticality – We define four categories for the number of affected connections in case of a power outage c_{ir} (1, 10, 100, 1000). We asked material specialists to categorize all items above €100.-. We could not obtain c_{ir} per region such that we assume that c_{ir} is equal over all regions. We estimate α_i, fraction of demand arising from a power outage for item i for all items at 0.1.

In Table 4.2 we use various measures of time, which are related as follows: 24 hour = 1 day, 7 days = 1 week, 52 weeks = 1 year.

Table 4.2
Parameter settings.

Parameter	Value	Remarks
LT_{i0}, LT_{i1}	0.5–16 weeks	
LT_{in}, $n\epsilon L_g$	3 days	
TE_{nr}, $n\epsilon L_4$	0.33 h	
TE_{nr}, $n\epsilon L_3$	0.33 or 0.75 h	0.75 when a manned warehouse delivers to a region corresponding to an unmanned warehouse in Network 4
TE_{1r}	1.5–2.5 h	
TE_{0r}, $TE_{\infty r}$	24, 72 h	
CE_{nr}, $n\epsilon\{\infty,0,1\}$	€205, €205, €80	
h_i	25%	% of unit item procurement price per year
FS_i	25%	As Liander has large suppliers, it is a conservative estimation
$\boldsymbol{\beta_{i0}}$	98%	Fixed fill rate for consignment stock
$\boldsymbol{\beta_{i1}}$	95%	Fill rate of *CW* for replenishment orders for low valued items

4.3. Results main model

4.3.1. Solution quality and computation time

We express the solution quality as the relative gap between the ILP solution value, TC_{ILP}, and the solution value of the LP-relaxation, TC_{LP}, i.e., $(TC_{ILP}\text{-}TC_{LP})/TC_{LP}$. The relative gap appears to be on average 0.035%, running 20 instances with *mx* values (target value for *MSPD*), between 0.1 and 2.0, optimizing 189 items. The maximum observed gap is 0.122%. As the optimal value of the LP-relaxation is a lower bound for the optimal ILP value, we conclude that the optimal ILP value is close to its lower bound. The average computation time is 2.2 min, with a maximum of 2.9 min on an AMD dual core 2.1 Ghz computer. The computation time is longer for low values of *mx*, as inventory levels are forced to be higher and more options should be considered. To gain more insight in the relative gap and calculation time we repeated the experiment on the same machine for 1000 items, where the additional items' characteristics are drawn from the original 189 items, with 20 *mx* values between 0.9 and 6.6. The relative gap improves to an average of 0.029% and a maximum of 0.095%. The computation time becomes on average 25.4 minutes with a maximum of 39.6 minutes.

4.3.2. Validation

We compared the stock allocation of the model with the current stock allocation. The stock allocation of the model is based on $mx=1$, as this seems to be a realistic target compared to the current performance of 2.07 *MSPD*. We observed that our model increases the assortment at a regional manned warehouse from the current 37–81 service parts types. The model removes 3 out of the current 37 items, each having no demand or being expensive (> €1200). Furthermore, we observed that the model roughly halves the value of these 34 such that the average value stored per item is lowered. 51% of the newly placed items have high impact ($c_{ir}=1000$), and 47% have a unit price below €200. So, we add cheap and extremely critical items to the regional stocks.

The proposed inventory allocation corresponds well with the perception of a regional warehouse manager. At the central warehouse, we increase the fill rate differentiation amongst items with a range between 51% and 100% instead of the current 90–98%. For the low valued items (cheaper than € 100), we could only obtain the current performance data for one regional warehouse in terms of the inventory costs. As the values for emergency shipment costs and *MSPD* are low in the model, we expect these values also to be low in the current situation. The model again halves the inventory value. However, we should note that in practice certain items are ordered by more than one at a time. Unfortunately, information on order size is unavailable for the low valued items.

4.3.3. Analysis cost versus MSPD

As it is hard for Liander to specify a clear target for the *MSPD*, we show the relation to the costs in Fig. 4.1. The current performance is also plotted in the graph. We see that costs sharply increase if *MSPD* drops below 1, so a good solution will be around that point. The numerical values for the combinations of total costs and *MSPD* on the efficient frontier, options, are given in Table 4.3.

We see that the solution corresponding to *MSPD* = 1, option 1, lowers the costs by 15% and the *MSPD* by 52%. Furthermore, it shows that if we pursue improvement solely in terms of costs (option 2) or *MSPD* (option 3), a reduction of 32% in costs or a reduction of 71% in *MSPD* is feasible. In the remainder of this study we use the solution corresponding to *MSPD* = 1.

For the low valued items, we had limited information from a single warehouse only. Optimizing 685 items at this regional warehouse resulted in a total costs of €9,761, where emergency shipment costs count for €41. The *MSPD* is 0.0009. The inventory costs of the current situation are €22,245, yielding a potential improvement of €12,484 at this warehouse. Extrapolation based on turnover of all regional warehouses yields a total improvement potential of €234,704 annually.

5. Simplifying the service part optimization

As Liander aims for a simple and practical methodology, we also develop an item approach proceeding from the results of the system approach. Section 5.1 discusses the rationale of the item approach and its method. For details, we refer to online Appendix E. Section 5.2 shows the costs of simplification by presenting the numerical results.

5.1. The item approach

5.1.1. Rationale

Obviously, a system approach yields the best solution, but the corresponding methodology is hard to grasp for practitioners. Therefore we approximate the benefits of a system approach by the simplicity of an item approach. Classification systems are developed in literature for single location systems, but to the best of our knowledge not for multi-echelon systems. Companies typically favor simple and transparent rules of thumb over complex optimization routines, which may come at the price of higher costs than strictly necessary. The added value in practice is that Liander does not need to rerun the total model for every new item which is added or removed (Hopp et al., 1999), and it provides insight in the relation between item characteristics and suitable network structures and waiting time. Also, we can estimate the costs of the simplification that Liander desires.

5.1.2. Method

The key idea is to find two statistical relations from our model results: (i) relating the network structure to item characteristics, (ii) relating the item waiting time to item characteristics. Next, we heuristically find the inventory allocation per item from these two statistical models. We briefly discuss these three steps. See Appendix E in the online supplement for a more comprehensive discussion.

5.1.2.1. Network structure. As the networks can be ranked from centralized to decentralized, the network number is an ordinal variable. We construct an ordinal logistic model to forecast the network structure per item, cf. Hosmer and Lemeshow, 2000. Fig. 5.1 shows a scatter plot of the network selected per item versus item price and demand * criticality, both on a logarithmic scale. The product of demand and criticality is selected, since it drives the left hand side of the *MSPD* constraint (6).

5.1.2.2. Target waiting time. We forecast the expected waiting time: $1/D_i \sum_{r \in R} EW_{ir}(\boldsymbol{p}_i) d_{ir}$ for item i over all regions. We apply a log-linear model, as we expect nonlinear relations between the waiting time and the explanatory variables (e.g., between mean demand and waiting time).

5.1.2.3. Inventory allocation. A final step is to find for each item the inventory allocation, given the network structure, the network with the highest probability, and the target waiting time from the regression models. We first correct for infeasible combinations of network structure and target waiting times. Next we apply different heuristics for the different network structures 2–4 which basically pursue to minimize the absolute deviation between the actual and the target waiting times. For network structure 3 and 4 the heuristic also aims for cost minimization of the inventory allocation.

5.2. Cost penalty of using the item approach

Table 5.1 compares the item approach to the results of the current and model performance.

We conclude that the item approach yields a costs increase of 10% and an *MSPD* increase of 28%, even though it is tuned to Liander's data. Three steps deteriorate the solution: (i) wrong network choice (ii) inaccurate target waiting time choice, (iii) inaccurate translation of the target waiting time into inventory levels. To find the impact of each step, we proceed as follows. We isolate the network forecast step by restricting the column

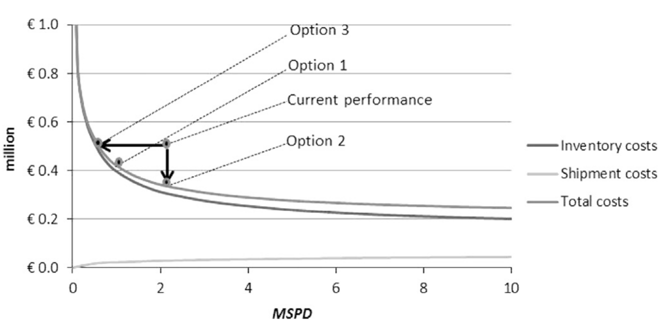

Fig. 4.1. Current performance and relation between costs and *MSPD* based on the parameter settings from Section 4.2.

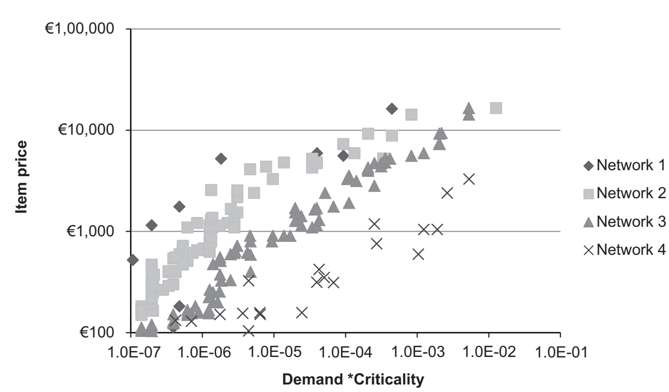

Fig. 5.1. Network selection depending on unit price and demand*criticality.

Table 4.3
Comparison of the current performance and the performance of different model allocations.

Item	Current	Option-1	Option-2	Option-3
Inventory costs (10^3)	€445	€398	€309	€478
Transport costs (10^3)	€54	€24	€31	€20
Total cost (10^3)	**€498**	€421	€340	**€498**
MSPD	**2.07**	1.00	**2.07**	0.602

Table 5.1
Comparison of the results with the item approach.

Item	Current	Model	Item approach
Inventory costs (10^3)	€445	€398	€432
Transport costs (10^3)	€54	€24	€28
Total cost (10^3)	€498	€421	€461
MSPD	2.07	1.00	1.28

Please cite this article as: van den Berg, D., et al., Allocating service parts in two-echelon networks at a utility company. International Journal of Production Economics (2015), http://dx.doi.org/10.1016/j.ijpe.2015.08.025

Table 5.2
Results for various assortments from model, item approach and an estimation of the current performance.

Experiment	Number of changed items	Model		Item approach		Estimated current	
		Cost(10^3)	*MSPD*	Cost(10^3)	*MSPD*	Cost(10^3)	*MSPD*
1	11	€479	1	€506	1.45	€567	2.07
2	18	€470	1	€495	1.43	€556	2.07
3	16	€435	1	€462	1.44	€515	2.07
4	18	€419	1	€448	1.36	€496	2.07
5	18	€449	1	€474	1.38	€531	2.07
6	15	€456	1	€486	1.41	€539	2.07
7	21	€445	1	€467	1.43	€526	2.07
8	*25*	*€604*	*1*	*€559*	*2.93*	*€714*	*2.07*
9	15	€432	1	€473	1.32	€511	2.07
10	12	€423	1	€465	1.28	€500	2.07
Average	**16.9**	**€461**	**1**	**€484**	**1.54**	**€546**	**2.07**

generation method of Section 3.4 to use the forecasted network. This approach yields an increase of costs and *MSPD* of 5% and 23%, respectively. Next, we only exclude the approximation of the target waiting times from the item approach by using the resulting waiting times from the model solution of Section 4.3 and apply the relevant methods of Section 5.1.2 to find the network and order parameters. We find nearly the same cost increase and an *MSPD* increase of 34%. Therefore, the performance loss is here mainly due to the regression results for the network and the translation of the target waiting time into stock levels.

In order to forecast the performance of the methodology in practice, we randomly remove items with a probability of 10%, and add the same number of new items, such that the number of items remains 189. The yearly turnover of assortment is 7% at Liander. The characteristics of the new items are randomly drawn from the characteristics of all 189 items. We observe that the costs of putting an item in Network 1 incorrectly can be large, therefore we overrule network choice 1 by 2 for new items. Recall that the regression model has difficulty forecasting Network 1 correctly. Further an allocation in Network 1 ignores the information of the target waiting time, as it only has one fill rate. On the other hand, Network 2 can differentiate fill rates between items, such that it is able to create a comparable allocation of Network 1. This ability is especially useful when the probabilities for Network 1 and 2 are close. Therefore, choosing Network 2 instead of Network 1 appears to be a more robust approach in our experiments. Table 5.2 compares the results of the item approach with the outcomes of the model and an estimation of the current performance. For the latter we assume that the same number of *MSPD* is caused as currently and that the costs increase in the same proportion as the exact model does between the original and the new assortment.

We observe that 9 out of 10 times the solution from the item approach dominates the solution of the estimated current situation. For the one that doesn't, experiment 8, a relatively good solution is found. The cost decreases by 28% but the *MSPD* increases by 29%. The *MSPD* performance on item level shows that in experiment 8, three items cause 45% of the *MSPD* in the item approach. By a manual adaptation of the network and target waiting time a better performance can easily be reached. So, the system should give a warning if one item causes more than a certain threshold *MSPD* (say 0.25 here), such that the user can manually adapt the targets for those items. Note that the *MSPD* caused is available on item level.

We see that our item approach improves the current practice: a costs decrease of 7%, and a decrease in *MSPD* by 38%. On average, without manual adaptations, these numbers are 11% and 26%. These are considerable improvements. Still, our system approach from Section 3 leads to better results, on average 5% in costs and 35% in *MSPD* compared to the item approach, and is also more generic (not dependent on case data). In practice, an item approach with simple guidelines for the service levels depending on item characteristics is often preferred over a system approach because of its transparency. Typically, this already leads to an improvement compared to a current situation in which service parts are not managed well. Here we observe the same phenomenon, but stress that even an advanced method to arrive at good item policies leads to solutions that are (significantly) dominated by the policies found using a system approach.

6. Conclusions

We show that by column generation and several single item models, with additional modifications and extensions, Liander is able to lower its costs while simultaneously decreasing its impact of waiting time for service parts by an improved stock allocation. We adapt the method of Özkan et al. (in press) such that lot sizing is possible at the CW and show by simulation that it can be used in practical settings. Further we develop an item approach, which aims to find a near-optimal solution to the two-echelon service part problem, resulting in a significant improvement compared to the current performance, 11% in costs and 26% in *MSPD* (Table 5.2, results 10 experiments). The simplicity of the item approach comes at a cost, despite the advanced method used, as the main model yields a significantly better solution, 5% in costs and 35% in *MSPD* (Table 5.2, results 10 experiments). That is the price the company has to pay for the desired simplicity.

Acknowledgments

The authors thank Liander for providing time and support and the anonymous reviewers for their helpful comments on the earlier version of this paper.

Appendix A. Procedure to find the steady state probabilities for Network 3 and 4 in Section 3.3.2

1. Start: $\hat{\pi}_{-\bar{S}} = 1$ (initial choice, in fact we express all other state probabilities in $\pi_{-\bar{S}}$)
   $d_1(-\bar{S}+1)\hat{\pi}_{-\bar{S}+1} = k(-\bar{S})\mu_1\hat{\pi}_{-\bar{S}} \Rightarrow \hat{\pi}_{-\bar{S}+1} = \frac{k(-\bar{S})\mu_1}{d_1(-\bar{S}+1)}\hat{\pi}_{-\bar{S}}$
2. Compute recursively: $\hat{\pi}_x = \frac{\mu_1}{d_1(x)}\sum_{y=max\{x-Q,-\bar{S}\}}^{x-1} k(y)\hat{\pi}_y$ for $x = -\bar{S}+2, -\bar{S}+3, \ldots, 0, 1, \ldots, s_1+Q_1$
3. Normalize the state probabilities: $\pi_x = \frac{\hat{\pi}_x}{\sum_{y=-\bar{S}}^{s_1+Q_1}\pi_y}$

Please cite this article as: van den Berg, D., et al., Allocating service parts in two-echelon networks at a utility company. International Journal of Production Economics (2015), http://dx.doi.org/10.1016/j.ijpe.2015.08.025

Appendix B. Algorithmic description of the adapted method of Özkan et al. (in press)

In order to give an algorithmic description we extend the notation of Section 3.2. For ease of notation we omit suffix i which refers to an item, suffix the demand by the regional warehouses and describe the algorithm for network 3. For network 4, L_3 should be replaced by L_4.

W_1 The mean delay for a replenishment order at the central warehouse.

RT_n The mean actual replenishment lead time of warehouse $n \epsilon \{2,\ldots,M+U\}$, $RT_n = LT_n + W_1$, where LT_n is the planned replenishment lead time of warehouse $n \epsilon \{0,\ldots,M+U\}$.

B_1 The mean number of backorders at the central warehouse.

An important logic of the method is that the fill rate, β_n, of the regional warehouses depends on the delay at the central warehouse, W_1. The delay at the central warehouse depends on the demand from the regional warehouses, which depends on the fill rate at the regional warehouses. The method of Özkan et al. (in press) solves this mutual dependence by an iterative procedure where β_n, $n \in L_3$ is calculated given W_1, step 2 and W_1 is calculated given β_n, $n \in L_3$, step 3. The method iterates over these steps until W_1 stabilizes, step 4.

1. Set $W_1 = 0$.
2. Compute β_n, $n \in L_3$.
   a. Determine RT_n by $RT_n = LT_n + W_1$, $n \in L_3$.
   b. Determine β_n by the Erlang loss formula $\beta_n = 1 - ERL(S_n, RT_n d_n)$, $n \in L_3$, Expression (8).

3. Compute W_1
   a. Determine the demand rate at the CW by:
   b. $d_1(x) = \{ \begin{matrix} \sum_{n \in L_3} d_n & if\ x > 0 \\ \sum_{n \in L_3} \beta_n d_n & if\ x \le 0 \end{matrix}$ Determine the steady state distribution of the CW inventory level, π_x, see Appendix A.
   c. Determine B_1 by $B_1 = \sum_{x=-\bar{S}}^{-1} (-x)\pi_x$.
   d. Determine W_1 from Little's formula by $W_1 = \frac{B_1}{\sum_{n \in L_3} \beta_n d_n}$

4. Iterate over step 2 and 3 until W_1 changes not more than ε.

5. Compute θ_n
   a. Determine β_1 by $\beta_1 = \sum_{x=1}^{s_1+Q_1} \pi_x$
   b. Determine θ_n by $\theta_n \approx \beta_1\ ERL(S_n, LT_n d_n)$, $n \in L_3$.
6. Compute γ_n by $\gamma_n = 1 - \beta_n - \theta_n$, $n \in L_3$.

The algorithmic description differs from the original method of Özkan et al. (in press) in Step 3b. The adaptation enables us to evaluate stock allocations where the replenishment order size at the central warehouse, Q_1, is larger than 1.

Appendix C. Alignment observations (Alvarez et al., in press) for column generation in Section 3.4.2.

We briefly discuss the adaptations we made to apply the method of Alvarez et al. (in press) as in our case: (i) Q_{i1} can be larger than 1, (ii) the difference in waiting time and cost calculation is dependent on three fractions instead of two and (iii) we consider two networks instead of one. The size of $Q_{i1} > 1$ allows for a tighter lower bound on the reduced costs of a new policy yielding a tighter upper bound on s_{i1}. The observation by Alvarez et al. (in press) yielding an upper bound on S_{in}, $n \epsilon \{2,\ldots,U+M\}$ notes that an increase of S_{in} only benefits the regions corresponding to that regional warehouse. Next, we find an upper bound on S_{in} when the holding costs increase of one additional item is higher than the resulting cost reduction in waiting time costs and emergency shipment costs. The maximum reduction in the waiting time for region r is in our case given by $\sum_{n \in L_g} LK_{nrg}(\theta_{in}(TE_{1r} - TE_{nr}) + \gamma_{in}(TE_{\infty r} - TE_{nr}))$. Finally, we have to consider two network types (3 and 4). We start with Network 3 in which a solution with minimal reduced costs is likely to be found the fastest, as it has the least number of regional warehouses. The policy found from Network 3 with corresponding reduced costs helps us to generate tighter bounds on stock levels for Network 4. The actual algorithm is somewhat different when the number of stock levels to evaluate at the CW is large (say, >15). To reduce the computation time, we then take bigger steps in the s_{i1} levels to evaluate, such that we choose to consider 5 stock levels between the lower and upper bound. Subsequently, we check if this results in a delivery policy with negative reduced costs. If this is not the case, all relevant stock levels at the CW will be evaluated. This saves computation time, since in the first number of iteration steps, the shadow prices do not accurately reflect the optimal balance between costs and waiting time.

Appendix D and E. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ijpe.2015.08.025.

References

Alfredsson, P., Verrijdt, J., 1999. Modeling emergency supply flexibility in a two-echelon inventory system. Manag. Sci. **45** (10), 1416–1431.

Alvarez, E.M., van der Heijden, M.C., 2014. On two-echelon inventory systems with Poisson demand and lost sales. Eur. J. Oper. Res. **235**, 334–338.

Alvarez, E.M., van der Heijden, M.C., Zijm, W.H.M., 2013. The selective use of emergency shipments for service-contract differentiation. Int. J. Prod. Econ. 143 (2), 518–526.

Alvarez, E.M., van der Heijden, M.C., Zijm, W.H.M., 2015. Service differentiation in spare parts supply through dedicated stocks. Ann. Oper. Res. (in press)

Andersson, J., Melchiors, P., 2001. A two-echelon inventory model with lost sales. Int. J. Prod. Econ. 69, 307–315.

Axsäter, S., 2006. Inventory Control, second edition. Springer science+Business Media, LLC, Berlin, Heidelberg.

Basten, R.J.I., van Houtum, G.J., 2014. System-oriented inventory models for spare parts. Surv. Oper. Res. Manag. Sci. 19, 34–55.

Bijvank, M., Vis, I.F.A., 2011. Lost-sales inventory theory: a review. Eur. J. Oper. Res. 215, 1–13.

Cohen, M., Kamesan, P.V., Kleindorfer, P., Lee, H., Tekerian, A., 1990. Optimizer: IBM's multi-echelon inventory system for managing service logistics. Interfaces 20, 65–82.

Hopp, W.J., Zhang, R.Q., Spearman, M.L., 1999. An easily implementable hierarchical heuristic for a two-echelon spare parts distribution system. IIE Trans. 31 (10), 977–988.

Özkan, E., van Houtum, G.J., Serin, Y., 2015. A new approximate evaluation method for two-echelon inventory systems with emergency shipments. Ann. Oper. Res., in press

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression, 2nd edition. Wiley, New York.

Korevaar, P., Schimpel, U., Boedi, R., 2007. Inventory budget optimization: meeting system-wide service levels in practice. IBM J. Res. Dev. 51 (3/4), 447–463.

Kranenburg, A.A., van Houtum, G.J., 2007. Effect of commonality on spare parts provisioning costs for capital goods. Int. J. Prod. Econ. 108, 221–227.

Kranenburg, A.A., van Houtum, G.J., 2008. Service differentiation in spare parts inventory management. J. Oper. Res. Soc. 59, 946–955.

Muckstadt, J.A., 2005. Analysis and Algorithms for Service Part Supply Chains. Springer, Berlin, Heidelberg..

Muckstadt, J.A., Thomas, L.J., 1980. Are multi-echelon inventory methods worth implementing in systems with low-demand-rate items? Manag. Sci. 26 (5), 483–494.

Rossetti, M.D., Achlerkar, A.V., 2011. Evaluation of segmentation techniques for inventory management in large scale multi-item inventory systems. Int. J. Logist. Syst. Manag. 8 (4), 403–424.

Şen, A., Bhatia, B., Doǧan, K., 2010. Applied materials uses operations research to design its service and parts network. Interfaces 40 (4), 253–266.

Sherbrooke, C.C., 1968. Metric: a multi-echelon technique for recoverable item control. Oper. Res. 16 (1), 122–141.

Sherbrooke, C.C., 2004. Optimal Inventory Modeling of Systems, second edition. Kluwer Academic Publishers, The Netherlands.

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 regional warehouse. OR Spectr. 29, 699–722.