European Journal of Operational Research 199 (2009) 908–921

Contents lists available at ScienceDirect

European Journal of Operational Research

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

A new partial pooling structure for spare parts networks ☆

A.A. Kranenburg [a,*], G.J. van Houtum [b,1]

[a] Centre for Quantitative Methods CQM B.V., P.O. Box 414, 5600 AK Eindhoven, The Netherlands
[b] School of Industrial Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 30 November 2007
Accepted 31 January 2009
Available online 31 March 2009

Keywords:
Inventory
Spare parts
Pooling
Lateral transshipment

ABSTRACT

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

1. Introduction

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

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

☆ This paper is based on Kranenburg (2006), for which the first author received the EURO Doctoral Dissertation Award 2007.

* Corresponding author. Tel.: +31 40 750 2322; fax: +31 40 750 1699.
E-mail addresses: kranenburg@cqm.nl (A.A. Kranenburg), g.j.v.houtum@tue.nl (G.J. van Houtum).
[1] Tel.: +31 40 247 5163; fax: +31 40 246 4596.

0377-2217/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2009.01.057

one of these local warehouses, a lateral transshipment will take place. This costs only about 10–20 hours time, and thus drastically reduces the waiting time for the spare part compared to an emergency shipment from the central warehouse. At ASML, in daily execution, lateral transshipment has been applied for years. However, in the *planning phase* in which base stock levels are determined, the option of lateral transshipment was *not* taken into consideration.

The contribution of this paper is as follows:

(i) *We introduce a network structure that is general and matches with structures observed in practice.* With respect to lateral transshipment, we distinguish two types of local warehouses in our network: *main* and *regular* local warehouses. While both main and regular local warehouses can receive a lateral transshipment, only main local warehouses are allowed to be the supplier of a lateral transshipment. Our network structure is interesting from a theoretical point of view. Consider the two extreme situations: on the one hand a situation with regular local warehouses only, and on the other hand a situation with main local warehouses only. The former constitutes the single-location situation without any lateral transshipment between local warehouses, or in other words, the *no pooling* situation. The latter case equals the situation with *full pooling*, i.e., where lateral transshipment is allowed between all local warehouses. Thus, *our network structure is a generalization* of which both the no pooling and full pooling situations are special cases. In general, our network structure leads to *partial pooling*, where only some of the local warehouses, namely the main local warehouses, are allowed to provide lateral transshipment. With our general network structure, we are able to compare no, partial and full pooling with respect to their performance and cost.
Our network structure also has a high practical relevance. In fact, the motivation for this network structure comes from the situation that we encountered at ASML: there exist differences between local warehouses. Some local warehouses are larger (have more inventory) since they serve larger or more customers and observe larger demand rates than other local warehouses. Some local warehouses have longer operating hours than others (for example also during the night). Further, some local warehouses are close to airports with many flights (hubs). It is desired that if a lateral transshipment is needed, it will be supplied by a local warehouse that can provide it quickly. In other words, especially local warehouses with the characteristics described above are suitable candidates to be main local warehouses. From the 19 local warehouses of ASML that are located in the United States of America, 4 local warehouses were identified as main local warehouses, for the above reasons. The other 15 local warehouses are regular local warehouses. Our network structure fits also for networks with so-called quick response stocks, which are being used by several companies; see, e.g. Rijk (2007). These companies do not allow lateral transshipment between real local warehouses, but they have installed extra warehouses from where parts can be delivered in case a local warehouse is out of stock. The real local warehouses may be modeled as regular local warehouses in our network structure, and the extra warehouses may be modeled as main local warehouses. In this case, the main local warehouses receive no direct demands from customers and their only function is to be supplier of lateral transshipment.

(ii) *We develop an approximate evaluation method for instances of real-life size.* The structure of our problem is such that the performance evaluation of the whole system with multiple SKU-s can be decomposed into an evaluation per SKU. Instances with a small number of local warehouses and small base stock levels can be analyzed exactly via a Markov process description. However, this does not work for instances of real-life size. Therefore we develop an approximate evaluation method that is based on the following main idea: (i) We decouple the network into individual local warehouses; (ii) We assume that the extra demand processes in main local warehouses due to requests for lateral transshipment are Poisson processes. Under the latter assumption, each individual warehouse can be analyzed analytically, and an iterative algorithm is needed to determine the demand rates for lateral transshipment. We show that the resulting approximate evaluation method is accurate and fast.

(iii) *We show that via partial pooling the major part of the full pooling benefits is obtained.* The approximate evaluation model is used in a greedy heuristic for the multi-item spare parts model with aggregate mean waiting time constraints. Based on this multi-item model, we show that partial pooling performs very well compared to full pooling. If only a few of the local warehouses are allowed to provide lateral transshipment, then already a substantial part of the full pooling benefits is obtained. This result has an important practical implication: only a few, well chosen local warehouses have to be equipped to provide lateral transshipment in order to obtain most of the benefits of full pooling.

(iv) *Our method leads to substantial cost savings for a case study with ASML data.* Previously, ASML used lateral transshipment between local warehouses in daily operation, but it did not take lateral transshipment into account when planning the base stock levels in the local warehouses. We show for data sets obtained from ASML that substantial savings (in the order of 30%) are realized when lateral transshipment is taken into account in the spare parts inventory planning method. Based on our case study, ASML decided to implement our greedy heuristic for their spare parts inventory model. They use our algorithm since early 2005 and succeeded to reduce both waiting times and cost substantially.

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

Wong et al. (2006) itself describe a continuous-review, multi-item spare parts system with two local warehouses with lateral transshipment and base stock control. Exact analysis of policies is done using a Markov process description. If the warehouse that faces a demand is out of stock, lateral transshipment from the other local warehouse is applied. If the other local warehouse has no stock on hand, an emergency replenishment from the central warehouse is carried out. It is assumed that the central warehouse has ample stock, so the model is a single-echelon model. Constraints are assumed on the aggregate mean waiting time, i.e., the average waiting time for an arbitrary request. Under these constraints, they minimize expected cost, consisting of holding and transportation cost for lateral and emergency shipments. To obtain a lower bound on the optimal cost for the multi-item problem, they use Lagrangian relaxation to separate the problem into multiple single-item optimization problems, and they use a sub-gradient method to find the optimal values for the Lagrange multipliers (there