A multi-item spare parts inventory model with customer differentiation
Kranenburg, A.A.; van Houtum, G.J.A.N.

Published: 01/01/2004

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the author's version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 27. Dec. 2018
A Multi-Item Spare Parts Inventory Model with Customer Differentiation

A.A. Kranenburg∗, G.J. van Houtum
Department of Technology Management, Technische Universiteit Eindhoven
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

February 2004

Abstract

We consider a single warehouse where spare parts of multiple stock-keeping units are kept on stock to serve customers with (close-to-)identical machines. Customers are divided into multiple customer classes, and a target aggregate fill rate is set per class. In order to get differentiated service levels, critical level policies are assumed. We formulate a multi-item, single-stage spare parts inventory model for this problem, with the objective to minimize the inventory investment under the condition that all target aggregate fill rates are met. We develop a solution procedure based on Lagrange relaxation, in order to obtain both a heuristic solution and a lower bound for the optimal costs. Underlying subproblems are solved by applying product-form solutions for closed queueing networks, an exact optimization procedure for a single-item spare parts inventory problem, and linear programming theory. An extensive computational experiment shows that the gap between the costs of the heuristic solution and the lower bound is small in general (on average 1.5%), and that computation times are limited (e.g., 3 minutes for instances with 20-100 items and 5 customer classes). The solution procedure is also applied to a case at ASML, a manufacturer of so-called step and scan systems, which are used for the production of integrated circuits. For their situation, applying critical level policies gives 10-20% reduction in inventory investment in comparison to simply using basestock policies (without critical levels) and providing the highest target aggregate fill rate to all customer classes. Although the focus of this paper is on the customer differentiation problem, our approach is described in general terms and seems applicable to several other multi-item spare parts problems.

Keywords: Inventory control, spare parts, system approach, multiple customer classes, customer differentiation, service level constraints, Lagrange relaxation.

∗Corresponding author. Phone: +31 40 2472637, Fax: +31 40 2464596, E-mail: a.a.kranenburg@tue.nl.
1 Introduction

The research in this paper is motivated by real-life spare parts networks for complex technical systems such as MRI scanners, large computer systems, airplanes, professional printing systems, and baggage systems. These systems are manufactured by Original Equipment Manufacturers (OEM-s) in small to moderate quantities and usually sold directly to customers (i.e., no intermediaries are involved). At the customers the availability of the installed technical systems often is essential for the primary processes. Hence, they require high availabilities. For that reason, first of all they have extensive preventive maintenance programs. Second, when components fail, they are replaced immediately by new or as-good-as-new components. If the failed part is repairable, the part is repaired off-line. The repair-by-replacement policy requires that new or ready-for-use spare parts are quickly available at the moments that failures occur. Since complex technical systems consist of thousands of components and many of them are subject to failures, spare parts of many different components have to be kept on stock. For individual companies it quickly becomes too expensive to keep all these spare parts on stock. Therefore, in several industries, the OEM has organized and controls the world-wide spare parts provisioning. He keeps spare parts on stock in a central stockpoint, at local stockpoints close to (groups of) customers, and possibly at intermediate stockpoints. Customer agreements are made on how quick spare parts have to be delivered and what is paid for that service by them; see e.g. Bol [2] and Kranenburg [13].

In the situation sketched above, the OEM is the central decision maker that controls the spare parts stocks in a divergent, multi-echelon inventory system. The objective of the OEM is to satisfy customer requirements with as low a total investment in spare parts as possible (for expensive systems, the investment in spare parts dominates total costs; transportation costs for regular replenishments and emergency and lateral shipments may be neglected). Customers may require a specified system availability, which is defined as the fraction of time that a system is not down because of a lack of spare parts. Or, alternatively, they may require a minimal aggregate fill rate level for the total number of demands for all components (each time that a demanded part cannot be delivered immediately their system goes down and their primary process is disturbed for some time; with emergency actions the consequences can be limited). Unders the presence of emergency deliveries, the two different service measures, system availability and aggregate fill rate, are coupled; see [13] and Section 5 of this paper. In either case, the service measure and service level requirements are system-oriented; i.e., they are not defined for individual components, but for the system as a whole. This requires a multi-item or system approach and constitutes a significant difference with standard multi-echelon inventory theory.

So, the OEM is faced with a multi-item, multi-echelon inventory problem in the above situation. In addition, the problem may incorporate a multi-indenture structure as well (if failed components
are repairable and subcomponents are needed for their repair). The main questions to be answered are: Which items have to be kept on stock, in which quantities and at which stockpoints of the distribution network, such that given service level (availability) constraints are met against a minimal total investment in spare parts? To answer this question, intelligent inventory control methods are needed. An existing theory with which optimal balances of inventories can be found is the METRIC theory. The METRIC theory consists of a whole series of multi-item models and corresponding solutions for spare parts distribution networks. We see the single-stage, multi-item model as studied by Feeney and Sherbrooke [9] as the first model in this area. The famous METRIC model itself is a two-echelon, multi-item model and was introduced and analyzed shortly later by Sherbrooke [18]. Further contributions have been made by several authors; for an overview, see e.g. Sherbrooke [19], and Rustenburg, Van Houtum and Zijm [16]. The METRIC theory has been applied in several (mainly military) environments, although often to a limited extent; see [5, 11, 19].

Considering the large savings that can be obtained by applying METRIC type models instead of straightforward inventory control models (think of savings of 20-50%; see e.g. [16, 19, 20]), the number of implementations in practice is disappointing. A reason might be that some essential properties such as emergency shipments, lateral transshipments, and customer differentiation, are not captured in the current models. The properties mentioned here are important issues at several companies, among which ASML with whom we collaborate. ASML is a leading manufacturer of so-called step and scan systems, which are used for the production of integrated circuits. We plan to develop METRIC type models in which these properties are incorporated. This paper is devoted to a first model that captures customer differentiation. The situation of ASML is used as guidance at points in the modelling and analysis where choices have to be made. ASML has centralized control of their complete spare parts network, consisting of one central warehouse and about 40 local warehouses at close distances to the factories of their customers (no spare parts are stocked at the factories in principle, and thus ASML has a two-echelon network). Service contracts are closed with all customers. The contracts include spare parts provisioning, for which each customer specifies either a target aggregate fill rate or a target system availability. System availability is one-to-one related to aggregate fill rate, but this relation depends on the distance between a factory where systems are installed and the supporting local warehouse (see Section 5). Hence, different target service levels may arise because of different service levels specified by customers (for the aggregate fill rate or the system availability) or because of differences in distances to a supporting local warehouse (when target system availabilities are specified).

The model studied in this paper is a single-stage, multi-item model with multiple customer classes and a target aggregate fill rate per customer class. Customers have similar machines and demands for spare parts occur according to Poisson demand processes. This model is applicable to each local warehouse in the situation of ASML as long as the central warehouse behaves as an ample
server, which is (almost) the case currently. The spare parts stocks are assumed to be controlled by a so-called critical level policy. This means that, per item, the total stock is controlled by a basestock policy and a critical level is specified per customer class. If a given customer class demands a part at a moment that the physical stock is at or below its critical level, then this demand will not be satisfied. Obviously, the higher the critical level of a customer class the lower the service level that is received. An unsatisfied demand is assumed to be satisfied by other channels (e.g. emergency shipments) and hence is a lost sale for the stock at the local warehouse. The objective is to find a critical level policy that minimizes the total inventory investment in spare parts subject to the aggregate fill rate constraints.

The minimization of the total inventory investment is a nonlinear integer optimization problem. Such problems are known to be hard, even for relatively small instances. We want to develop a solution procedure that is applicable to large problems (several customer classes and 1000 items, say). Therefore, it is impossible to develop an appropriate exact solution procedure, and thus we have to be satisfied with a heuristic solution procedure. It appears that, due to the structure of our problem, an efficient heuristic solution procedure based on Lagrange relaxation can be developed. In the approach, or analysis, leading to this procedure, we use product-form solutions for closed queueing networks, an exact optimization procedure for single-item problems, and linear programming theory. The approach seems quite generic; i.e., we believe that the approach is applicable to several other multi-item spare parts problems (see the beginning of Section 3 where the approach is outlined in quite general terms). Our solution procedure generates both a heuristic solution and a lower bound for the optimal costs. An extensive computational experiment consisting of instances with 20-100 items and 2-5 customer classes shows that the procedure is both accurate and relatively fast. In this experiment, the relative distance between the cost of the heuristic solution and the lower bound is 1.5% on average. Also, this relative distance decreases strongly as a function of the number of items (see Section 4), and thus a high accuracy may be expected for real-life problems with many items. The computation time was equal to 9 seconds on average for instances with 20-100 items and 2 customer classes, and equal to 3 minutes for instances with 20-100 items and 5 customer classes.

So far, customer differentiation has been studied in the literature for single-item, single-stage cases only. The problem has been introduced by Veinott [22]. He also introduced the concept of critical level policies. Further, we distinguish studies where the structure of optimal policies is investigated and studies where evaluation and optimization within a given class of policies is solved. Within the first stream, there are interesting studies that derive the optimality of critical level policies for single-item models with multiple customer classes. Topkis [21] considers a periodic-review model with generally distributed demand and zero leadtime. In that situation, the optimal critical levels are dependent on the remaining time in a period. Ha [12] has studied a continuous-
review model with Poisson demand processes, a single exponential server for replenishments, and lost sales. He derives the optimality of critical level policies, and in this situation both the basestock levels and critical levels are time-independent (Ha's results seem extendable to the case with an ample server). De Véricourt, Karaesmen, and Dallery [3] studied the same model as Ha but with backordering of unsatisfied demand instead of lost sales, and obtained the same results. Within the second stream of studies, there are interesting contributions by Dekker, Hill, Kleijn, and Teunter [4], Melchiors, Dekker, and Klein [14], and Deshpande, Cohen, and Donohue [6]. Dekker et al. [4] derive exact procedures for the generation of an optimal critical level policy for a continuous-review model with multiple customer classes, Poisson demands, ample supply, and lost sales. Melchiors et al. [14] extend this work to a model with fixed quantity ordering and two customer classes. In this model, the fixed ordering size, the basestock level, and a single critical level policy are optimized in order to minimize the sum of fixed ordering, inventory holding, and lost sales costs. Deshpande et al. [6] consider a similar model but with backordering of unsatisfied demand. In that situation one also must decide in which order backordered demands are satisfied, which leads to additional complications. For further contributions, see the references in the above papers.

For our model, we limit ourselves to the class of critical level policies with time-independent critical and basestock levels. The results on optimal policies for single-item problems support our choice to limit ourselves to critical level policies. Further, by the results of Ha [12] and De Véricourt et al. [3], it is appropriate to assume time-independent parameters in case of exponential replenishment leadtimes. In case of non-exponential leadtimes, optimal parameters are likely to become dependent on the status of the outstanding orders (cf. Topkis [21]). However, the latter would be too complicated for practical purposes, and thus time-independent parameters seem also fair to assume for non-exponential replenishment leadtimes.

The main contribution of this paper is as follows. First, this paper is the first study on a multi-item spare parts inventory model with customer differentiation. Second, we derive a fast and accurate heuristic solution procedure. Third, the approach leading to the solution procedure seems applicable to several other multi-item problems. Fourth, we apply our solution procedure to data for one representative local warehouse of ASML, and compare costs under our heuristic solution to the costs that are obtained when using pure basestock policies and providing the highest target fill rate to all customer classes (also called round-up policy by Deshpande et al. [6]). The computational results shows that, by the use of critical level policies, ASML can realize a reduction in total inventory investment at the local warehouses by 10-20% (which corresponds to tens of millions of EURO-s in absolute terms).

This paper is organized as follows. The model is described in Section 2, and the approach leading to the efficient solution procedure is given in Section 3. The computational experiment is executed in Section 4, and the case study is presented in Section 5. After that, extensions are
discussed in Section 6. Finally, conclusions and possible directions for further research are given in Section 7.

2 Model

Consider a single warehouse that keeps several spare parts on stock to serve a number of customers. Customers have factories where (close-to-)identical machines are installed. Machines generate demands for spare parts, and we assume that the rates per machine are (close-to-)identical. We distinguish $I \ (\in \mathbb{N})$ stock-keeping units (SKU-s), numbered $i = 1, \ldots, I$, and $J \ (\geq 2)$ customer classes, numbered $j = 1, \ldots, J$. (We assume that $J \geq 2$, but the solution procedure that we obtain ultimately also applies to the case with one customer class.)

If one of the parts of a machine fails, the machine goes down and the defective part has to be replaced by a ready-for-use part to get the system up and running again. This ready-for-use part is demanded from the warehouse. A failure of a machine is always caused by one defective part and can be remedied by replacing that part only. For each SKU $i$ and customer class $j$, failures (demands) are assumed to occur according to a Poisson process with a constant rate $m_{i,j} \ (\geq 0)$. This assumption is common in METRIC type models. In reality, the failure rate is lower for a while when a machine is down, but this effect is negligible if down times are short (relative to the time between two successive failures). Also, if there are many machines within a customer class, then the decrease in the failure rate is relatively small anyway. We define $M_j$ as the total demand for customer class $j$, i.e., $M_j := \sum_{i=1}^{I} m_{i,j}$, and we assume that $M_j > 0$, $j = 1, \ldots, J$.

For each customer class $j$, a target aggregate fill rate $\beta_{j,\text{obj}}$ has been specified. This target aggregate fill rate is the minimum percentage of demands of customer class $j$ that has to be delivered immediately upon request. W.l.o.g., we assume that customer class are ranked in order of nonincreasing targets, i.e., $\beta_{1,\text{obj}} \geq \ldots \geq \beta_{J,\text{obj}}$. (Recall that aggregate fill rate is a system-oriented service measure, and that there is a coupling with system availability.)

The spare parts stocks are assumed to be controlled by a continuous-review critical level policy. This means that, for each SKU $i$, the total stock (i.e., the inventory position) is controlled by a basestock policy, with basestock level $c_{i,j+1} \ (\in \mathbb{N}_0 := \{0\} \cup \mathbb{N})$ and there is a critical level $c_{i,j} \ (\in \mathbb{N}_0)$ per customer class $j = 1, \ldots, J$. If customer class $j$ demands a part at a moment that the physical stock of SKU $i$ is larger than $c_{i,j}$, then this demand is satisfied and otherwise not. In the latter case, the demand is fulfilled by another channel, e.g. by an emergency shipment from either the supplier or another source. For the stock at the warehouse such a demand is a lost sale. The larger $c_{i,j}$ the lower the service level that is received. In the extreme case with $c_{i,j} = 0$, demands of class $j$ are fulfilled immediately as long as there is physical stock available at the warehouse. In the other extreme case with $c_{i,j}$ larger than or equal to the basestock level, no demand of class $j$ is
fulfilled immediately. The combination of a basestock policy and continuous review implies one-for-
one replenishments. Each satisfied demand triggers a replenishment of a ready-for-use part. The
replenishment leadtimes for an SKU are assumed to be independent and identically distributed,
and leadtimes for different SKU-s are independent. The mean replenishment leadtime of SKU \( i \) is
\( t_i \ (> 0) \). The underlying process may have different forms. In case of both consumable SKU-s and
repairable SKU-s, the underlying process may be a regular replenishment process where the supplies
come from a supporting/central warehouse within the same organization (like in the situation of
ASML when the model is applied to a local warehouse; at ASML failed parts of repairable SKU-s
are sent into repair by the central warehouse only), or the supplies come from an external supplier.
Alternatively, in case of repairable SKU-s, the underlying process may be a repair process at an
internal or external repair facility with ample capacity. Another possibility is a combination that
is obtained for a repairable SKU subject to condemnation. Then each failed part can be repaired
with a specified probability. If so, then the failed part is sent into repair to replenish the stock and
otherwise a new part is procured (the mean \( t_i \) then is a weighted mean of both channels).

The critical level policy for SKU \( i \) is denoted by both \((c_{i,1}, \ldots, c_{i,J}, c_{i,J+1})\) and \( \tilde{c}_i := (\bar{c}_i, c_{i,J+1}) \)
with \( \bar{c}_i := (c_{i,1}, \ldots, c_{i,J}) \). A customer class \( j > 1 \) has a target aggregate fill rate that is at most
equal to the target for any class \( k < j \), and thus it seems reasonable to give a critical level to class
\( j \) that is at least equal to the critical level of class \( k \). I.e., we assume that \( c_{i,1} \leq c_{i,2} \leq \ldots \leq c_{i,J} \).
Further, a positive value is given to a critical level \( c_{i,j} \) if we want to limit the access of a customer
class \( j \) in favor of more important classes. For class 1 there is no more important class and thus we
set \( c_{i,1} = 0 \) (if \( c_{i,1} > 0 \), then \( c_{i,1} \) parts of SKU \( i \) would never be accessible and thus would constitute
dead stock). Finally, for any customer class \( j \), the access to the stock of SKU \( i \) is completely
blocked if \( c_{i,j} \geq c_{i,J+1} \). W.l.o.g., we may exclude that \( c_{i,j} > c_{i,J+1} \). So, for each SKU \( i \), we have
\( 0 = c_{i,1} \leq c_{i,2} \leq \ldots \leq c_{i,J} \leq c_{i,J+1} \). A (complete) critical level policy denotes the policy for all
SKU-s, and is denoted by \( c := (c_1, \ldots, c_I) \).

The objective is to minimize the total inventory investment subject to the aggregate fill rate
constraints for the customer classes. Because the demand processes of all SKU-s are independent
Poisson processes, the aggregate fill rate of customer class \( j \), \( \beta_j(c) \), is a weighted sum of fill rates
for individual SKU-s, with the fractions \( m_{i,j}/M_j \) as weights:

\[
\beta_j(c) = \sum_{i=1}^{I} \frac{m_{i,j}}{M_j} \beta_{i,j}(c_i), \quad j = 1, \ldots, J,
\]

where \( \beta_{i,j}(c_i) \) denotes the item fill rate received by customer class \( j \) for SKU \( i \). Expressions for
these item fill rates are derived in the next section. For the inventory investment, we distinguish
two possibilities. If the parts in the pipeline (i.e., on order) are included, then the total inventory
investment is given by \( \sum_{i=1}^{I} p_i c_{i,J+1} \), where \( p_i > 0 \) denotes the price of SKU \( i \). This is appropriate
in case the parts are supplied by a supporting warehouse within the same organization and in case all
parts are repairables. For SKU-s supplied by an external supplier, it is more appropriate to exclude the pipeline stock. Then the inventory investment for an SKU $i$ becomes equal to $p_i(c_{i,J+1} - \sum_{j=1}^J m_{i,j}\beta_{i,j}(c_i)t_i)$, and this term replaces the term $p_i c_{i,J+1}$. In the situation of ASML, our model is applicable to a local warehouse and local warehouses are supplied by a central warehouse. Therefore we choose to include the pipeline stock for all SKU-s. However, the solution procedure that we develop is easily adapted for the case where pipeline stock has to be excluded for one or more SKU-s; see Subsection 6.1.

In mathematical terms, our optimization problem is as follows:

$$(P) \quad \text{min} \quad \sum_{i=1}^I p_i c_{i,J+1}$$

subject to

$$\sum_{i=1}^I \frac{m_{i,j}}{M_j} \beta_{i,j}(c_i) \geq \beta_{j,\text{obj}}, \quad j = 1, \ldots, J,$$

$$0 = c_{i,1} \leq \ldots \leq c_{i,J} \leq c_{i,J+1}, \quad i = 1, \ldots, I,$$

$$c_{i,1}, \ldots, c_{i,J}, c_{i,J+1} \in \mathbb{N}_0, \quad i = 1, \ldots, I.$$

The optimal costs of Problem $(P)$ are denoted by $C_P$. Problem $(P)$ has a linear objective function, nonlinear constraints, and integral decision variables. It thus is a nonlinear integer programming problem.

Finally, at the end of this modelling section, we like to return to the assumption that machines and the corresponding failure rates are identical, or at least close-to-identical, for different customers. This assumption assures that (about) the same aggregate fill rate is obtained by customers within the same customer class (they then have the same item fill rates and the same weights in the expression for their aggregate fill rates; see (1)). Thus the targets of all customers are met if the target for the whole class is met. Also, in the computational experiment in Section 4, we limit ourselves to instances for which this assumption is satisfied. For the rest, the assumption is not used; see also Subsection 6.2.

3 Analysis

Because Problem $(P)$ is nonlinear and integral, it is a hard problem to solve exactly, even for relatively small instances. Our focus is on instances with several customer classes and many SKU-s (500, say), and therefore we have to be satisfied with a good heuristic solution. We will develop a fast and accurate heuristic solution procedure based on Lagrange relaxation (see Porteus [15], Appendix B, for a good description of this technique applied to general constrained optimization problems). The solution procedure leads to both a heuristic solution and a lower bound for the optimal costs.
The approach leading to the heuristic solution procedure consists of the following steps. In the first preliminary step, based on the theory of closed product-form networks, closed-form formulae are derived for the item fill rates, by which exact evaluation of a given critical level policy is obtained. Next, Lagrange relaxation is applied. After that, the optimization for given Lagrange multipliers is considered. Due to the separability of the objective function and the aggregate fill rates, this problem decomposes into single-item problems for which an exact solution can be found efficiently. Subsequently, the Lagrange parameters are optimized in order to obtain the best possible lower bound for the optimal costs. Due to the integrality of the solutions space (i.e., the integrality of the basestock and critical levels), this problem can be transformed into a linear programming problem that is solvable by an efficient method. Finally, a heuristic solution is constituted by one of the optimal critical level policies under the optimized Lagrange parameters, or by slight modifications of one of these policies. In total we have 5 steps, which are described in the Subsections 3.1-3.5, respectively.

### 3.1 Evaluation of a critical level policy

In this subsection, we derive closed-form formulae for the item fill rates $\beta_{i,j}(c_i)$, and by that we have exact closed-form formulae for the evaluation of a given critical level policy $c$.

For an SKU $i$, the number of parts in the replenishment pipeline plus the number of parts in the physical stock is always equal to the basestock level $c_{i,J+1}$. A part remains in the pipeline according to a given replenishment leadtime distribution with mean $t_i$, and this time is independent of other parts that may be present in the pipeline. After leaving the pipeline, the part is added to the physical stock. Parts in the physical stock leave the stock when demands are fulfilled by the physical stock. W.l.o.g., we may assume that the oldest part is taken when a demand is fulfilled (i.e., the parts leave the physical stock in FCFS order). Because of the exponential demand processes, we have an exponential time until the next demand occurs, but, because of the role of the critical levels, the rate depends on the size of the physical stock. If the number of parts in the physical stock is $k \in \{0, \ldots, c_{i,J+1}\}$, then the rate equals $\mu_{i,k} = m_{i,1} + \ldots + m_{i,j}$ where $j$ is the largest index for which $c_{i,j} < k$ (for $k = 0$, we have $\mu_{i,0} = 0$). When a demand is fulfilled, the number of parts in the physical stock decreases with 1 and the number in the pipeline is increases with 1.

From the description above, it follows that the behavior of SKU $i$ is described by a closed queueing network with $c_{i,J+1}$ customers and two stations: (i) an ample server with mean service time $t_i$, which represents the pipeline stock; (ii) a load-dependent, exponential, single server with FCFS service discipline, which represents the physical stock. The service rates of the load-dependent server are given by the $\mu_{i,k}$. This network belongs to the class of so-called BCMP networks and thus has a product-form solution; see Baskett et al. [1]. By applying the theory of [1], we find that the steady state probability $q_{i,k}$ for having $k$ parts in the pipeline, and thus $c_{i,J+1} - k$ parts in the
physical stock, equals
\[ q_{i,k} = \left\{ \prod_{h=0}^{k-1} \tilde{\mu}_{i,h} \right\} \frac{t_{i,k}}{k!} q_{i,0}, \quad k = 1, \ldots, c_i, J+1, \quad (2) \]

with
\[ q_{i,0} = \left( \sum_{k=0}^{c_i,J+1} \left\{ \prod_{h=0}^{k-1} \tilde{\mu}_{i,h} \right\} \frac{t_{i,k}}{k!} \right)^{-1}, \]

\[ \tilde{\mu}_{i,h} = \mu_{i,c_i,J+1-h} \text{ for } h \in \{0, \ldots, c_i,J+1\}, \]

and the convention that \( \prod_{h=0}^{k-1} \tilde{\mu}_{i,h} = 1 \) for \( k = 0 \) (this result also follows from Gnedenko and Kovalenko [10], pp. 250–252). By these steady state probabilities, we obtain the item fill rates:
\[ \beta_{i,j}(c_i) = \sum_{k=0}^{c_i,J+1-c_i,j-1} q_{i,k}, \quad j = 1, \ldots, J, \]

with the convention that this sum is empty when \( c_i,J+1-c_i,j-1 < 0 \) (i.e., \( \beta_{i,j}(c_i) = 0 \) if \( c_i,j = c_i,J+1 \)). Notice that \( 1 \geq \beta_{i,1}(c_i) \geq \ldots \geq \beta_{i,j}(c_i) \geq 0. \)

### 3.2 Lagrange relaxation

We now apply Lagrange relaxation to Problem \((P)\) (cf. Porteus [15], Appendix B). By this relaxation, the \( J \) service level constraints of Problem \((P)\) are replaced by \( J \) additional terms in the objective function. To create the Lagrangian problem, we define \( J \) Lagrangian multipliers. To smooth the analysis, we use the products \( \lambda_j M_j, j = 1, \ldots, J, \) for these multipliers, where the \( \lambda_j \) are nonnegative, real-valued parameters. This results in the following problem for a given vector \( \lambda, \) with \( \lambda := (\lambda_1, \ldots, \lambda_J): \)

\[ (LR(\lambda)) \min \sum_{i=1}^{I} p_i c_i,J+1 + \sum_{j=1}^{J} \sum_{i=1}^{I} \lambda_j m_{i,j} (\beta_{j,\text{obj}} - \beta_{i,j}(c_i)) \]

subject to \( 0 = c_{i,1} \leq \ldots \leq c_{i,J} \leq c_{i,J+1}, \quad i = 1, \ldots, I; \)
\[ c_{i,1}, \ldots, c_{i,J}, c_{i,J+1} \in \mathbb{N}_0, \quad i = 1, \ldots, I. \]

Note that the term \( m_{i,j} (\beta_{j,\text{obj}} - \beta_{i,j}(c_i)) = m_{i,j} (1 - \beta_{i,j}(c_i)) + m_{i,j} (\beta_{j,\text{obj}} - 1) \) can be interpreted as the expected number of lost sales per time unit for item \( i \) and customer class \( j, \) plus a constant term \( m_{i,j} (\beta_{j,\text{obj}} - 1). \) Hence, the parameter \( \lambda_j \) can be interpreted as the cost of a lost sale for customer class \( j. \) The objective function of Problem \((LR(\lambda))\) thus consists of the inventory investment, penalty costs for all demands that are not fulfilled immediately, and a constant term.

Let \( c_{LR(\lambda)}^* \) denote an optimal policy for Problem \((LR(\lambda)), \) and let \( C_{LR(\lambda)} \) denote the corresponding optimal costs. Furthermore, let Problem \((LR)\) be defined as

\[ (LR) \max C_{LR(\lambda)} \]

subject to \( \lambda_j \geq 0, \quad j = 1, \ldots, J. \)
This problem may be denoted as a dual problem to Problem \((P)\) (cf. [15]). Let \(\lambda^*_LR\) and \(C_{LR}\) denote an optimal solution of this problem and the corresponding optimal costs, respectively. Under an optimal solution \(\lambda^*_LR = (\lambda^*_LR,1,\ldots,\lambda^*_LR,J)\), the penalty cost parameters \(\lambda^*_LR,j\) are balanced such that under an optimal critical level policy of the corresponding Problem \((LR(\lambda^*_LR))\) the aggregate fill rates of the customer classes \(j = 1,\ldots,J\) are as close to the target aggregate fill rates as possible. (A \(\lambda^*_LR,j\) may be equal to 0 in which case the aggregate fill rate of class \(j\) under an optimal policy of \((LR(\lambda^*_LR))\) may be significantly larger than the target for class \(j\); this may occur when the aggregate fill rate constraint for class \(j\) is not binding in Problem \((P)\), which in general is unlikely for our problem.)

Obviously, for any nonnegative \(\lambda\), Problem \((LR(\lambda))\) provides a lower bound on the optimal solution of Problem \((P)\) (see e.g. [15], p. 250), and the tightest bound is obtained by Problem \((LR)\), i.e.,

\[
C_P \geq C_{LR}.
\]

The difference \(C_P - C_{LR}\) is known as the duality gap. In our problem, we have an integral solution space for Problem \((P)\) and therefore the gap is positive in general (only for special instances, we obtain a zero gap; see also the following remark).

**Remark 1 (on the Everett Result)** Let \(\lambda = (\lambda_1,\ldots,\lambda_J) \geq 0\) and let \(c^*_LR(\lambda)\) be an optimal policy for Problem \((LR(\lambda))\). Under this policy, the aggregate fill rate for customer class \(j\) is equal to \(\beta_j(c^*_LR(\lambda)), j = 1,\ldots,J\). By a result of Everett [8] (see also [15], p. 245), the policy \(c^*_LR(\lambda)\) is optimal for Problem \((P)\) for any target fill rates \(\beta_j,obj\) for which the following two conditions are satisfied:

\[
\beta_j,obj \leq \beta_j(c^*_LR(\lambda)), \quad j = 1,\ldots,J
\]

\[
\sum_{j=1}^J \lambda_j (\beta_j(c^*_LR(\lambda)) - \beta_j,obj) = 0.
\]

If \(\lambda_j > 0\) for all \(j\), then these two conditions reduce to

\[
\beta_j,obj = \beta_j(c^*_LR(\lambda)), \quad j = 1,\ldots,J.
\]

This result offers an alternative for solving Problem \((P)\). The Everett result says that for any \(\lambda\) (with positive components \(\lambda_j\), say) and a corresponding optimal policy \(c^*_LR(\lambda)\) for Problem \((LR(\lambda))\), the policy \(c^*_LR(\lambda)\) is also optimal for Problem \((P)\) with target aggregate fill rates \(\beta_j(c^*_LR(\lambda))\). By generating optimal policies for vectors \(\lambda\) under which the \(\beta_j(c^*_LR(\lambda))\) are close to the \(\beta_j,obj\), one obtains a series of optimal solutions for slightly different targets than the \(\beta_j,obj\). This is an appropriate alternative for real-life problems. In fact, this method constitutes an analogue of the greedy procedures used for METRIC-type problems by which a series of efficient solutions is obtained for the inventory investment and one given service measure; see [19] and [16].
3.3 Optimization for given Lagrange multipliers

In this subsection, we derive an exact solution for problem \((LR(\lambda))\), where \(\lambda\) is an arbitrary vector with nonnegative components.

The objective function of Problem \((LR(\lambda))\) can be rewritten as

\[
\sum_{i=1}^{I} \left\{ p_i c_{i,J+1} + \sum_{j=1}^{J} \lambda_j m_{i,j} \ (1 - \beta_{i,j}(c_i)) \right\} - \sum_{j=1}^{J} \lambda_j M_j \ (1 - \beta_{j,\text{obj}}). \tag{6}
\]

The term \(\sum_{j=1}^{J} \lambda_j M_j \ (1 - \beta_{j,\text{obj}})\) is constant for given values \(\lambda_j\) and thus may be ignored for the optimization. Then the remaining problem decomposes into \(I\) independent single-item minimization problems, i.e., into one problem per SKU \(i\):

\[
(SI(i)) \quad \min \quad p_i c_{i,J+1} + \sum_{j=1}^{J} \lambda_j m_{i,j} \ (1 - \beta_{i,j}(c_i))
\]

subject to \(0 = c_{i,1} \leq \ldots \leq c_{i,J} \leq c_{i,J+1},\)

\[c_{i,1}, \ldots, c_{i,J}, c_{i,J+1} \in \mathbb{N}_0.\]

For this problem, an efficient solution procedure is obtained based on a convex lower bound function. The lower bound function is derived along the same lines as for a single-item problem studied by Dekker et al. [4]. The convexity is proved by applying a result of Dowdy et al. [7].

First, we distinguish solutions for Problem \((SI(i))\) on the basis of the basestock level \(c_{i,J+1}\). We define \(f_i(c_{i,J+1})\) as the minimal value that is obtained for Problem \((SI(i))\) under a fixed value \(c_{i,J+1} \in \mathbb{N}_0\):

\[
f_i(c_{i,J+1}) = \min \quad p_i c_{i,J+1} + \sum_{j=1}^{J} \lambda_j m_{i,j} \ (1 - \beta_{i,j}(\tilde{c}_i, c_{i,J+1}))
\]

subject to \(0 = c_{i,1} \leq \ldots \leq c_{i,J} \leq c_{i,J+1},\)

\[c_{i,1}, \ldots, c_{i,J} \in \mathbb{N}_0.\]

For a given \(c_{i,J+1}\), \(f_i(c_{i,J+1})\) can be found by explicit enumeration over all feasible values of \(\tilde{c}_i\). An optimal policy \((\tilde{c}_i, c_{i,J+1})\) corresponding to \(f_i(c_{i,J+1})\) is denoted as \((\tilde{c}_i^*(c_{i,J+1}), c_{i,J+1})\). Then Problem \((SI(i))\) is equivalent to minimizing function \(f_i(c_{i,J+1})\), i.e., to

\[
(SI(i))' \quad \min \quad f_i(c_{i,J+1})
\]

subject to \(c_{i,J+1} \in \mathbb{N}_0.\)

Let \(\lambda_{\text{min}} := \min \lambda_j\). We introduce a function \(g_i(c_{i,J+1}), c_{i,J+1} \in \mathbb{N}_0\), identical to \(f_i(c_{i,J+1})\) but
with the penalty costs $\lambda_j$ for all $j$ decreased to $\lambda_{\text{min}}$:

$$g_i(c_{i,J+1}) = \min_p p_i c_{i,J+1} + \lambda_{\text{min}} \sum_{j=1}^{J} m_{i,j} (1 - \beta_{i,j}(\tilde{c}_i, c_{i,J+1}))$$

subject to $0 = c_{i,1} \leq \ldots \leq c_{i,J} \leq c_{i,J+1}$,

$$c_{i,1}, \ldots, c_{i,J} \in \mathbb{N}_0.$$  

This latter optimization problem corresponds to a single-item problem with multiple customer classes and the same cost of a lost sale per class. Then there is no incentive to differentiate between customer classes, and thus the costs are minimized by $\tilde{c}_i = 0$. This implies that $\beta_{i,j}(\tilde{c}_i, c_{i,J+1}) = \beta_{i,j}(0, c_{i,J+1})$ for all $j$. Hence, $g_i(c_{i,J+1})$ may be rewritten as

$$g_i(c_{i,J+1}) = p_i c_{i,J+1} + \lambda_{\text{min}} \sum_{j=1}^{J} m_{i,j} (1 - \beta_{i,j}(0, c_{i,J+1})).$$

For function $g_i(c_{i,J+1})$, we obtain the following two lemmas.

**Lemma 1** $g_i(c_{i,J+1}) \leq f_i(c_{i,J+1})$ for all $c_{i,J+1} \in \mathbb{N}_0$.

**Proof:** Let $c_{i,J+1} \in \mathbb{N}_0$. For any feasible $\tilde{c}_i$, the objective function in the formulation of $g_i(c_{i,J+1})$ is smaller than or equal to the objective function in the formulation of $f_i(c_{i,J+1})$, and thus $g_i(c_{i,J+1}) \leq f_i(c_{i,J+1})$. $\square$

**Lemma 2** $g_i(c_{i,J+1})$ is convex, i.e., $g_i(c_{i,J+1} + 2) - 2g_i(c_{i,J+1} + 1) + g_i(c_{i,J+1}) \geq 0$, $c_{i,J+1} \in \mathbb{N}_0$.

**Proof:** To show that function $g_i(c_{i,J+1})$ is convex, it suffices to show that $\beta_{i,1}(0, c_{i,J+1})$ is concave as a function of $c_{i,J+1} \in \mathbb{N}_0$.

As we showed in Subsection 3.1, the behavior of SKU $i$ is described by a closed queueing network with $c_{i,J+1}$ customers and two stations: (i) an ample server with mean service time $t_i$; (ii) an exponential, single server. Under policy $(0, c_{i,J+1})$, all critical levels are zero and thus the single server becomes load-independent. We are interested in the throughput of the network. Because the throughput is independent of the distribution of the service times at the ample server, we may assume w.l.o.g. that the service times at the ample server are exponential. The ample server then may be seen as a load-dependent, exponential, single server, with a service rate that is linear with the number of jobs. Hence, we can apply a result of Dowdy et al. [7]. They prove that the throughput in a closed queueing network with one load-independent, exponential, single server and one load-dependent, exponential, single server, whose load-dependent service rates are nondecreasing and concave (of which our linearly increasing service rate is a special case), is concave as a function
of the number of customers in the network. It is easily verified that in our network the throughput is equal to \( \beta_{i,j}(0, c_{i,J+1}) \sum_{j=1}^{J} m_{i,j} \). The property that \( \beta_{i,j}(0, c_{i,J+1}) \sum_{j=1}^{J} m_{i,j} \) is concave in \( c_{i,J+1} \) implies that \( \beta_{i,j}(0, c_{i,J+1}) \) is concave as a function of \( c_{i,J+1} \) \( \in \mathbb{N}_0 \).

Let \( c_{i,J+1}^*(x) := \arg \min \{ f_i(c_{i,J+1}) | c_{i,J+1} = 0, \ldots, x \} \), i.e., \( c_{i,J+1}^*(x) \) denotes the optimal value for \( f_i(c_{i,J+1}) \) up to point \( x \).

**Lemma 3** If \( f_i(c_{i,J+1}^*(x)) \leq g_i(x + 1) \) and \( g_i(x) \leq g_i(x + 1) \) for some \( x \in \mathbb{N}_0 \), then critical level policy \((\check{c}_i^\ast(c_{i,J+1}^*(x)), c_{i,J+1}^*(x)) \) is an optimal solution of Problem \((SI(i)')\) and \( f_i(c_{i,J+1}^*(x)) \) is equal to the corresponding optimal costs.

**Proof:** Lemmas 1 and 2 imply that if we find a value \( x \) for which \( g_i(x) \leq g_i(x + 1) \), then \( f_i(c_{i,J+1}) \geq g_i(x + 1) \) for all \( c_{i,J+1} \geq x + 1 \). If in addition \( f_i(c_{i,J+1}^*(x)) \leq g_i(x + 1) \), then function \( f_i(c_{i,J+1}) \) is minimal at \( c_{i,J+1}^*(x) \). Hence, \((\check{c}_i^\ast(c_{i,J+1}^*(x)), c_{i,J+1}^*(x)) \) is an optimal solution of Problem \((SI(i)')\), and the optimal costs of Problem \((SI(i)')\) are given by \( f_i(c_{i,J+1}^*(x)) \).

Based on Lemma 3, we obtain the following exact solution procedure for Problem \((SI(i)')\), or equivalently, for Problem \((SI(i))\). Notice that during execution, \((\check{c}_i^\ast(c_{i,J+1}^*(x)), c_{i,J+1}^*(x)) \) contains the best solution obtained so far. At termination, \((\check{c}_i^\ast(c_{i,J+1}^*(x)), c_{i,J+1}^*(x)) \) contains an optimal solution.

**Algorithm 1:** Exact algorithm for Problem \((SI(i))\)

**Initialization:**

- \( x := 0 \)
- \((\check{c}_i^\ast(c_{i,J+1}^*(x)), c_{i,J+1}^*(x)) := (0, 0)\)

**While not** \((f_i(c_{i,J+1}^*(x)) \leq g_i(x + 1)) \) and \((g_i(x) \leq g_i(x + 1)) \) **do**
- \( x := x + 1 \)
- If \( f_i(x) < f_i(c_{i,J+1}^*(x)) \), then \((\check{c}_i^\ast(c_{i,J+1}^*(x)), c_{i,J+1}^*(x)) := (\check{c}_i^*(x), x)\)

**End While**

Finally, \( c_{LR(\lambda)}^* = ((\check{c}_1^*(c_{1,J+1}^*(x)), c_{1,J+1}^*(x)), \ldots, (\check{c}_I^*(c_{I,J+1}^*(x)), c_{I,J+1}^*(x))) \), and \( C_{LR(\lambda)} \) can be found by plugging \( c_{LR(\lambda)}^* \) into (6).

The complexity of this algorithm is linear with the number of feasible solutions \((\check{c}_i, c_{i,J+1}) \) with \( c_{i,J+1} \leq x \), where \( x \) is the smallest value for which \( f_i(c_{i,J+1}^*(x)) \leq g_i(x) \) and \( g_i(x) \leq g_i(x + 1) \). For
the problems we are interested in, demand rates are (very) small and thus the optimal basestock level and this \( x \) are expected to be small as well. Hence the number of solutions that has to be considered will be small.

### 3.4 Optimization of the Lagrange multipliers

In this subsection, we derive a solution procedure for Problem (\( LR \)). We first transform Problem (\( LR \)) into a linear programming problem.

Problem (\( LR \)) may be re-written as

\[
\begin{align*}
\text{max} \quad & y \\
\text{subject to} \quad & y \leq C_{LR(\lambda)} \\
& \lambda_j \geq 0, \quad j = 1, \ldots, J.
\end{align*}
\]

Define \( C \) as the set of all critical level policies. The elements of \( C \) are denoted by \( c^k = (c^k_1, \ldots, c^k_{J+1}) \), where \( c^k_i = (c^k_{i,1}, \ldots, c^k_{i,J+1}), \quad 0 = c^k_{i,1} \leq \ldots \leq c^k_{i,J} \leq c^k_{i,J+1}, \quad c^k_{i,1}, \ldots, c^k_{i,J+1} \in \mathbb{N}_0, \quad i = 1, \ldots, I. \) Notice that the number of elements of \( C \) is infinite. For each critical level policy \( c^k \), we define \( w^k := \sum_{i=1}^I p_i c^k_{i,J+1}, \) and \( v^k := (v^k_1, \ldots, v^k_J) \), with \( v^k_j := \sum_{i=1}^I m_{i,j} (\beta_{j,\text{obj}} - \beta_{i,j}(c^k_i)) \). Then Problem (\( LR(\lambda) \)) may be rewritten as

\[
\begin{align*}
\text{max} \quad & w^k + \sum_{j=1}^J v^k_j \lambda_j \\
\text{subject to} \quad & c^k \in C.
\end{align*}
\]

Hence, the first constraint in Problem (\( LR' \)) may be replaced by

\[
y \leq w^k + \sum_{j=1}^J v^k_j \lambda_j, \quad \forall c^k \in C.
\]

We then obtain the following problem that is equivalent to Problem (\( LR \)):

\[
\begin{align*}
\text{(LRC(\( C \)))} \quad \text{max} \quad & y \\
\text{subject to} \quad & y \leq w^k + \sum_{j=1}^J v^k_j \lambda_j, \quad \forall c^k \in C, \\
& \lambda_j \geq 0, \quad j = 1, \ldots, J.
\end{align*}
\]

More generally, for each subset \( \mathcal{C} \subseteq C \), we define Problem (\( LRC(\mathcal{C}) \)) identical to Problem (\( LRC(\mathcal{C}) \)) defined above, but with \( C \) replaced by \( \mathcal{C} \). I.e., in Problem (\( LRC(\mathcal{C}) \)), we do not take into account all constraints given by (7), but only a subset of them.
Let $(\lambda_{LRC}(\hat{C}), y_{LRC}(\hat{C}))$ denote an optimal solution of Problem $(LRC(\hat{C}))$, and let the corresponding optimal costs be denoted by $C_{LRC}(\hat{C}) = y_{LRC}(\hat{C})$. Notice that $\lambda_{LRC}(C) = \lambda^{*}_{LR}$ and $C_{LRC}(C) = C_{LR}$. For each $\hat{C} \subseteq C$, $C_{LRC}(\hat{C})$ constitutes an upper bound for $C_{LRC}(C)$, and the larger $\hat{C}$ the tighter this bound. For each finite subset $\hat{C}$, Problem $(LRC(\hat{C}))$ is a straightforward linear programming problem in $\lambda_{j}$, $j = 1, \ldots, J$ that can be solved by the simplex method; see e.g. Schrijver [17] for a description of the simplex method. We use Problem $(LRC(\hat{C}))$ to solve Problem $(LRC(C)) = (LR)$ as follows.

First, we solve Problem $(LRC(\hat{C}))$ for a small initial set of constraints $\hat{C}$. Then, for the resulting solution $\lambda_{LRC}(\hat{C})$, we determine a constraint (i.e., a critical level policy) that has been exceeded with the largest surplus, i.e., we solve Problem $(LR(\lambda))$ for $\lambda = \lambda_{LRC}(\hat{C})$. We add this constraint to $\hat{C}$ and solve Problem $(LRC(\hat{C}))$ again. This process is repeated until there is no constraint that cuts off the current solution of Problem $(LRC(\hat{C}))$. We then know that $C_{LRC}(\hat{C}) = C_{LR}$; see the following lemma.

**Lemma 4** If $C_{LRC}(\hat{C}) = C_{LR}(\lambda_{LRC}(\hat{C}))$, then $C_{LR} = C_{LRC}(\hat{C})$.

**Proof:** By definition, $C_{LR}(\lambda) \leq C_{LR}$ for any $\lambda \geq 0$, and obviously, $C_{LR} \leq C_{LRC}(\hat{C})$ for any $\hat{C} \subseteq C$. Hence, if $C_{LRC}(\hat{C}) = C_{LR}(\lambda_{LRC}(\hat{C}))$, then $C_{LRC}(\hat{C}) = C_{LR}(\lambda_{LRC}(\hat{C})) \leq C_{LR}$, which implies $C_{LR} = C_{LRC}(\hat{C})$. □

The condition $C_{LRC}(\hat{C}) = C_{LR}(\lambda_{LRC}(\hat{C}))$ is used as stopping criterion for our procedure. To start the procedure, we need a finite initial set $\hat{C}$ for which a finite optimal solution $\lambda_{LRC}(\hat{C})$ is obtained. To specify such a set, we need the following definition. A set of $K = J + 1$ constraints

$$y \leq u^{K} + \sum_{j=1}^{J} u^{j} \lambda_{j}, \quad k = 1, \ldots, K,$$

is said to be a basis if the $K$ vectors $u^{k}$ have the following properties:

1. $v^{1}, \ldots, v^{K}$ span $\mathbb{R}^{J}$;

2. there exist $K$ nonnegative numbers $u_{k}$ such that $\sum_{k=1}^{K} u_{k} v^{k} = 0$, with $(u_{1}, \ldots, u_{k}) \neq (0, \ldots, 0)$.

Problem $(LRC(\hat{C}))$ can be solved by the simplex method to a finite optimal solution if and only if a basis exists in Problem $(LRC(\hat{C}))$. Define $S_{i} := \min\{c_{i,J+1}|\beta_{i,1}(0,c_{i,J+1}) > \beta_{1,0}\}$, $i = 1, \ldots, I$. Let $c^{1} := ((0, \ldots, 0), \ldots, (0, \ldots, 0))$, $c^{2} := ((0, \ldots, 0, S_{1}), \ldots, (0, \ldots, 0, S_{J}))$, $c^{3} := ((0, \ldots, 0, S_{1}, S_{1}), \ldots, (0, \ldots, 0, S_{I}, S_{I}))$, $\ldots$, $c^{K} := ((0, S_{1}, \ldots, S_{1}), \ldots, (0, S_{I}, \ldots, S_{I}))$. It may be verified that $\hat{C}_{ini} := \{c^{1}, \ldots, c^{K}\}$ forms a basis. We use this set as the initial subset, and obtain the following exact solution procedure for Problem $(LR)$.
Algorithm 2: Exact algorithm for Problem (LR)

Initialization:

- $\hat{C} := \hat{C}_{ini}$
- Solve Problem ($LRC(\hat{C})$) with the simplex method

While $C_{LR(\lambda_{LRC}(\hat{C}))} < C_{LRC(\hat{C})}$ do

- $\hat{C} := \hat{C} \cup \{c^*_k\}$
- Solve Problem ($LRC(\hat{C})$) with the (dual) simplex method

End While

That this algorithm is finite is seen as follows. Under the initial subset $\hat{C}_{ini}$, we obtain a finite value $C_{LRC(\hat{C}_{ini})}$. In each iteration step, the constraint that is added to $\hat{C}$ has to be binding in Problem ($LRC(\hat{C})$); i.e., it must be a policy $c^k$ for which $w^k + \sum_{j=1}^{J} v_j^k \lambda_j \leq C_{LRC(\hat{C})}$ and thus also $w^k \leq C_{LRC(\hat{C}_{ini})}$. The number of policies $c^k$ for which $w^k \leq C_{LRC(\hat{C}_{ini})}$ is finite, and hence the number of iteration steps is finite too.

3.5 Heuristic solution

Finally, in this subsection, we described how a feasible solution for Problem ($P$) is obtained.

At termination of Algorithm 2, the binding constraints form a basis. It can, but not necessarily will, happen that one of these constraints is a critical level policy that satisfies all $J$ service level constraints in Problem ($P$), and that we thus have obtained a feasible solution for Problem ($P$). In case no feasible policy is present in the basis, we apply local search (a greedy procedure). From the $K = J + 1$ policies in the basis, we choose the policy $c^k$ with the smallest distance

$$D := \sum_{j=1}^{J} \left( \beta_{j, obj} - \beta_{j}(c^k) \right)^+, $$

where $[x]^+ := \max\{0, x\}$. Starting with this policy, we iteratively evaluate a number of neighboring policies and select the best one, until we obtain a feasible solution for Problem ($P$). A policy $c^{k'}$ is a neighbor of policy $c^k$ if all critical and basestock levels are identical, except for one SKU $i$, for which $c^{k'}_{i,J+1} = c^k_{i,J+1} + 1$. Notice that this implies that $\beta_{j}(c^{k'}) \geq \beta_{j}(c^k), j = 1, \ldots, J$. Neighboring policies are evaluated with respect to the decrease in distance per unit cost,

$$\frac{1}{p_i} \left( \sum_{j=1}^{J} \left[ \beta_{j, obj} - \beta_{j}(c^k) \right]^+ - \sum_{j=1}^{J} \left[ \beta_{j, obj} - \beta_{j}(c^{k'}) \right]^+ \right),$$

and the neighbor for which this value is largest is selected.
4 Computational Experiment

In this section, we execute a computational experiment to test our solution procedure. In this experiment, we compare the costs of the heuristic solution, which constitute an upper bound ($UB$) for the optimal costs $CP$ of Problem ($P$), to the lower bound ($LB$) constituted by $CLR$. We define the relative distance $G$ between both bounds as

$$G := \frac{UB - LB}{LB}.$$  

We take instances with 2 and 5 customer classes, and different settings for the other parameters. For all instances, we log how the upper bound is obtained (i.e., as a binding constraint at termination of Algorithm 2 or via local search). The choice of the values for demand rates, target fill rates, ranges of prices, and mean replenishment leadtime is partly based on what we observed in the situation of ASML.

Table 1 contains the parameter settings for the instances with $J = 2$ customer classes, and Table 2 for the instances with $J = 5$. In the experiment we assume identical machines, which implies that parameters $m_{i,j}$ are fully determined by the total demand rate per SKU and the ratio between the demand rates for different customer classes. For the prices $p_i$ and the total demand rates $\sum_j m_{i,j}$ per SKU, uniform distributions are specified from which the $p_i$ and $\sum_j m_{i,j}$ are drawn (per SKU $i$). Because Problem ($P$) is an integer programming problem, the values of $LB$, $UB$, and $G$ significantly depend on the precise values of demand rates and other input variables. So, there is a certain level of coincidence involved. To exclude the effect of this factor on mean values for $G$, for both $J = 2$ and $J = 5$, we generated 20 samples for each combination of the number of SKU-s, the uniform distribution used for the prices of SKU-s, and the uniform distribution used for the total demand rates per SKU. This gives already 240 instances for both $J = 2$ and $J = 5$. In addition, for $J = 2$, for each of these instances, we varied the target aggregate fill rates and the ratio of the demand rates for both customer classes. As a result, we obtain 2880 instances in total with $J = 2$ customer classes, 240 instances with $J = 5$, and 3120 instances all together.

In Tables 3 to 11, we display the results of the experiment. In each table, we list: (i) the (rounded) average value of $LB$ for specified values of the input parameters; (ii) the percentage of instances where local search ($LS$) had to be applied to obtain a feasible solution, together with the average number of iterations in that case ($NumIt$); (iii) the average value of $G$; (iv) the maximum value of $G$. The model has been implemented in AIMMS 3.3, with CONOPT used to solve the LP problems, and we used a Pentium 4 computer. For the situation with 2 customer classes, average and maximum computation time per run were 9 and 42 seconds, respectively. In the situation with 5 customer classes, the average computation time was 201 seconds, whereas the maximum computation time approached 14 minutes.

Concerning the relative distance $G$ between the lower and upper bound, we observe that $G$
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Number of choices</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SKU-s (I)</td>
<td>3</td>
<td>20, 50, 100</td>
</tr>
<tr>
<td>Target aggregate fill rates ($\beta_{j,\text{obj}}$)</td>
<td>4</td>
<td>(0.99, 0.95), (0.99, 0.90), (0.98, 0.95), (0.98, 0.90)</td>
</tr>
<tr>
<td>Price ($p_i$)</td>
<td>2</td>
<td>U[1, 100], U[1, 10000]</td>
</tr>
<tr>
<td>Total daily demand rate per SKU</td>
<td>2</td>
<td>U[0, 0.01], U[0, 0.1]</td>
</tr>
<tr>
<td>Ratio ($m_{i,1} : \ldots : m_{i,2}$)</td>
<td>3</td>
<td>(0.2 : 0.8), (0.5 : 0.5), (0.8 : 0.2)</td>
</tr>
<tr>
<td>Mean repl. leadtime ($t_i$; in days)</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1: Parameter settings for the situation with 2 customer classes

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Number of choices</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of SKU-s (I)</td>
<td>3</td>
<td>20, 50, 100</td>
</tr>
<tr>
<td>Target aggregate fill rates ($\beta_{j,\text{obj}}$)</td>
<td>1</td>
<td>(0.99, 0.98, 0.95, 0.90, 0.80)</td>
</tr>
<tr>
<td>Price ($p_i$)</td>
<td>2</td>
<td>U[1, 100], U[1, 10000]</td>
</tr>
<tr>
<td>Total daily demand rate per SKU</td>
<td>2</td>
<td>U[0, 0.01], U[0, 0.1]</td>
</tr>
<tr>
<td>Ratio ($m_{i,1} : \ldots : m_{i,5}$)</td>
<td>1</td>
<td>(0.2 : \ldots : 0.2)</td>
</tr>
<tr>
<td>Mean repl. leadtime ($t_i$; in days)</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2: Parameter settings for the situation with 5 customer classes

<table>
<thead>
<tr>
<th></th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LB_{av}$ (*1000)</td>
<td>83</td>
<td>218</td>
<td>435</td>
</tr>
<tr>
<td>$LS$; NumIt</td>
<td>24%; 2.3</td>
<td>28%; 2.6</td>
<td>30%; 3.2</td>
</tr>
<tr>
<td>$G_{av}$ (%)</td>
<td>2.76</td>
<td>1.05</td>
<td>0.52</td>
</tr>
<tr>
<td>$G_{max}$ (%)</td>
<td>14.21</td>
<td>4.04</td>
<td>1.89</td>
</tr>
</tbody>
</table>

Table 3: Results for 2 customer classes and varying values for $I$

<table>
<thead>
<tr>
<th></th>
<th>(0.99, 0.95)</th>
<th>(0.99, 0.90)</th>
<th>(0.98, 0.95)</th>
<th>(0.98, 0.90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LB_{av}$ (*1000)</td>
<td>265</td>
<td>255</td>
<td>234</td>
<td>226</td>
</tr>
<tr>
<td>$LS$; NumIt</td>
<td>28%; 2.8</td>
<td>21%; 2.6</td>
<td>35%; 2.8</td>
<td>26%; 2.8</td>
</tr>
<tr>
<td>$G_{av}$ (%)</td>
<td>1.43</td>
<td>1.27</td>
<td>1.46</td>
<td>1.62</td>
</tr>
<tr>
<td>$G_{max}$ (%)</td>
<td>8.48</td>
<td>9.03</td>
<td>9.41</td>
<td>14.21</td>
</tr>
</tbody>
</table>

Table 4: Results for 2 customer classes and varying values for $\beta_{j,\text{obj}}$

...decreases significantly if the number of SKU-s $I$ increases (see Tables 3 and 8). This is an important observation because in real-life situations in which this model will be applied, the number of SKU-s...
<table>
<thead>
<tr>
<th>LBav (*1000)</th>
<th>U[1, 100]</th>
<th>U[1, 1000]</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS; NumIt</td>
<td>26%; 2.7</td>
<td>29%; 2.8</td>
</tr>
<tr>
<td>Gav (%)</td>
<td>1.40</td>
<td>1.49</td>
</tr>
<tr>
<td>Gmax (%)</td>
<td>14.21</td>
<td>11.34</td>
</tr>
</tbody>
</table>

Table 5: Results for 2 customer classes and different uniform distributions for the prices $p_i$

<table>
<thead>
<tr>
<th>LBav (*1000)</th>
<th>U[0, 0.01]</th>
<th>U[0, 0.1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS; NumIt</td>
<td>33%; 2.9</td>
<td>22%; 2.5</td>
</tr>
<tr>
<td>Gav (%)</td>
<td>1.96</td>
<td>0.93</td>
</tr>
<tr>
<td>Gmax (%)</td>
<td>14.21</td>
<td>5.93</td>
</tr>
</tbody>
</table>

Table 6: Results for 2 customer classes and different uniform distributions for the total demand rates per SKU

<table>
<thead>
<tr>
<th>LBav (*1000)</th>
<th>(0.2 : 0.8)</th>
<th>(0.5 : 0.5)</th>
<th>(0.8 : 0.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS; NumIt</td>
<td>237</td>
<td>245</td>
<td>253</td>
</tr>
<tr>
<td>Gav (%)</td>
<td>1.71</td>
<td>1.43</td>
<td>1.19</td>
</tr>
<tr>
<td>Gmax (%)</td>
<td>14.21</td>
<td>11.83</td>
<td>10.10</td>
</tr>
</tbody>
</table>

Table 7: Results for 2 customer classes and varying choices for the ratios $m_{i,1} : m_{i,2}$

<table>
<thead>
<tr>
<th>LBav (*1000)</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS; NumIt</td>
<td>95</td>
<td>226</td>
<td>439</td>
</tr>
<tr>
<td>Gav (%)</td>
<td>2.93</td>
<td>1.22</td>
<td>0.71</td>
</tr>
<tr>
<td>Gmax (%)</td>
<td>6.96</td>
<td>4.16</td>
<td>1.76</td>
</tr>
</tbody>
</table>

Table 8: Results for 5 customer classes and varying values for $I$

<table>
<thead>
<tr>
<th>LBav (*1000)</th>
<th>U[1, 100]</th>
<th>U[1, 10000]</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS; NumIt</td>
<td>46%; 2.3</td>
<td>45%; 2.6</td>
</tr>
<tr>
<td>Gav (%)</td>
<td>1.62</td>
<td>1.62</td>
</tr>
<tr>
<td>Gmax (%)</td>
<td>6.20</td>
<td>6.96</td>
</tr>
</tbody>
</table>

Table 9: Results for 5 customer classes and different uniform distributions for the prices $p_i$
Table 10: Results for 5 customer classes and different uniform distributions for the total demand rate per SKU

<table>
<thead>
<tr>
<th></th>
<th>U[0, 0.01]</th>
<th>U[0, 0.1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LB_{av}$ (*1000)</td>
<td>184</td>
<td>322</td>
</tr>
<tr>
<td>$LS; NumIt$</td>
<td>44%; 2.6</td>
<td>47%; 2.4</td>
</tr>
<tr>
<td>$G_{av}$ (%)</td>
<td>2.03</td>
<td>1.21</td>
</tr>
<tr>
<td>$G_{max}$ (%)</td>
<td>6.96</td>
<td>5.10</td>
</tr>
</tbody>
</table>

Table 11: Results for complete experiment

<table>
<thead>
<tr>
<th></th>
<th>$J = 2$</th>
<th>$J = 5$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LB_{av}$ (*1000)</td>
<td>245</td>
<td>253</td>
<td>246</td>
</tr>
<tr>
<td>$LS; NumIt$</td>
<td>27%; 2.8</td>
<td>45%; 2.5</td>
<td>29%; 2.7</td>
</tr>
<tr>
<td>$G_{av}$ (%)</td>
<td>1.45</td>
<td>1.62</td>
<td>1.46</td>
</tr>
<tr>
<td>$G_{max}$ (%)</td>
<td>14.21</td>
<td>6.96</td>
<td>14.21</td>
</tr>
</tbody>
</table>

usually will be large. Furthermore, higher values of the total demand rates result in smaller values of $G$ (see Tables 6 and 10), and $G$ is rather insensitive for the other inputs that have been varied. In general, $G$ is small (on average 1.46%), which implies that the obtained heuristic solution has a reasonably high accuracy. For the maximum gap $G_{max}$, a high value has been obtained for the 960 instances with $I = 20$ and $J = 2$, while much lower values were found for instances with 50 and 100 SKU-s (see Table 3). Apparently, with unlucky values of the input parameters for instances with a small number of SKU-s one can get a much bigger gap than for instances with many more SKU-s. This is also observed for the instances with $J = 5$ customer classes (see Table 8).

From Table 11, it can be seen that the percentage of instances where local search had to be applied to obtain a feasible solution was much higher for the instances with 5 customer classes than for the instances with 2 customer classes. Further, this percentage appears to be sensitive for the difference in target aggregate fill rates (see Table 4), the total demand rates (see Table 6 for the instances with $J = 2$; this sensitivity is not found for the instances with $J = 5$, see Table 9), and for the sizes of the customer classes (see Table 7). In total, in 29% of all instances, local search has been applied. For these 900 instances (29% of 3120), $G_{av} = 1.63\%$, while $G_{av} = 1.39\%$ for the 2220 instances where no additional local search was required at termination of Algorithm 2. This relatively small difference makes us think that improving the local search method could only lead to minor improvements.
5 Case Study: ASML

In this section, we present a case study at ASML. Recall that ASML's spare parts network consists of one central warehouse and about 40 local warehouses. In this case study, we consider the spare parts control for the two main machine families at one representative local warehouse. We compare the costs obtained under customer differentiation via critical level policies to the costs obtained by simply using basestock policies and providing the highest target aggregate fill rate to all customer classes. The latter is also denoted as the round-up policy (cf. Deshpande et al. [6]).

Our model allows us to define different target fill rate levels for different customer classes. In practice, however, some of ASML's customers specify target service levels in terms of time that machines (at a customer) are down due to the lack of spare parts. This measure is called 'down time waiting parts' and reflects directly unavailability of machines. In the scientific literature, this down time is also referred to as time that a machine is logistically unavailable. Under the presence of emergency shipments, or another channel, that is used to satisfy a demand in case a demand is not fulfilled from the stock in the local warehouse, there is a one-to-one relationship between down time for a customer class and its aggregate fill rate at the warehouse. Let $DT_{j, obj} \geq 0$ denote the maximum allowed total down time (per time unit) for customer class $j$, $T_t \geq 0$ the transportation time from the warehouse to the customer, and $T_e \geq 0$ the time needed to satisfy a demand by an emergency shipment. Then this relation is as follows:

$$DT_{j, obj} = M_j \left[ T_t + (1 - \beta_{j, obj})T_e \right].$$

Notice that even if $\beta_{j, obj}$ would be equal to 1, then a non-zero down time is implied if $T_t > 0$. In the case of ASML, representative values are 0.5 to 2 hours for $T_t$ and 6 to 36 hours for $T_e$. For example, if $T_t = 40$ minutes, $T_e = 36$ hours, $M_j = 0.4$ per week, then a target maximum down time $DT_{j, obj}$ of 0.5 hours per week implies a target fill rate level $\beta_{j, obj}$ of 0.984.

The first machine family that we consider, is the STEPPER family, which consists of more or less identical machines, with in total 352 relevant spare parts (i.e., spare parts with a positive failure rate for at least one customer class). We consider 5 customer classes, with target aggregate fill rates 0.99, 0.97, 0.95, 0.93, and 0.90, respectively. These are realistic values, partly based on an underlying maximum down time. The failure rates per SKU for all customer classes together vary from 0.003 to 35.1 per year. The average yearly failure rate is 1.41, and thus very small. The regular replenishment leadtime equals 7 days for all SKU-s (i.e., $t_i = 7$ days for all $i$). For the STEPPER family, we compare different options in Table 12. As a starting point, the round-up policy has been used, i.e. with all 5 customer classes together considered as one class having a target aggregate fill rate of 0.99. The costs of the obtained feasible solution has been normalized to 1000. To evaluate the round-up policy, we used the same algorithm as for the case with $J \geq 2$ classes ($J \geq 2$ is what we assumed in our model description and analysis; it appears that our solution
procedure reduces to an appropriate procedure for the case with \( J = 1 \). Besides the situation with 5 customer classes, situations with 2 customer classes have been considered, where in both classes the highest aggregate fill rate of all subclasses is required. These 2 classes can be formed in 4 different sensible ways. In Table 12, \( LB, UB \), and \( G \) are shown for each situation, as well as the number of SKU-s with non-zero critical level(s) for the obtained heuristic solution. Table 13 contains similar results for a second machine family at ASML, the STEP & SCAN family, for which again 5 customer classes are considered, with target fill rate levels of 0.99, 0.95, 0.90, 0.85, and 0.80, respectively. In the STEP & SCAN family, we have 465 spare parts with yearly failure rates per SKU for all customer classes together varying from 0.006 to 117, and with an average of 2.28. Again, we have \( t_i = 7 \) days for all \( i \). For both machine families, the range of prices of the SKU-s is large. The ratio between the most expensive and cheapest item is in the order of \( 10^5 - 10^6 \) (if the 10% most expensive SKU-s and 10% cheapest are excluded, then this ratio reduces to \( 10^2 \)). The computation time was equal to around 200 minutes for all 12 cases/instances together (140 minutes were needed for the case with 5 customer classes for the STEPPER family).

<table>
<thead>
<tr>
<th>Number of classes</th>
<th>Composition</th>
<th>( LB )</th>
<th>( UB )</th>
<th>( G ) (%)</th>
<th>Number of SKU-s with non-zero critical level(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>988</td>
<td>1000</td>
<td>1.19</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1 - 2345</td>
<td>928</td>
<td>930</td>
<td>0.22</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>12 - 345</td>
<td>921</td>
<td>922</td>
<td>0.07</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>123 - 45</td>
<td>936</td>
<td>938</td>
<td>0.14</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>1234 - 5</td>
<td>926</td>
<td>928</td>
<td>0.21</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>897</td>
<td>901</td>
<td>0.42</td>
<td>57</td>
</tr>
</tbody>
</table>

Table 12: Normalized results for STEPPER family

<table>
<thead>
<tr>
<th>Number of classes</th>
<th>Composition</th>
<th>( LB )</th>
<th>( UB )</th>
<th>( G ) (%)</th>
<th>Number of SKU-s with non-zero critical level(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>1000</td>
<td>1000</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1 - 2345</td>
<td>852</td>
<td>855</td>
<td>0.36</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>12 - 345</td>
<td>828</td>
<td>837</td>
<td>1.10</td>
<td>106</td>
</tr>
<tr>
<td></td>
<td>123 - 45</td>
<td>868</td>
<td>869</td>
<td>0.13</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>1234 - 5</td>
<td>932</td>
<td>944</td>
<td>1.23</td>
<td>279</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>800</td>
<td>801</td>
<td>0.11</td>
<td>230</td>
</tr>
</tbody>
</table>

Table 13: Normalized results for STEP & SCAN family

On average, \( G \) is 0.40% for the 10 cases with multiple customer classes in this case study, and in
4 cases local search was needed to obtain a feasible solution. Furthermore, the obtained basestock levels $c_{i,J+1}$ are low (mostly 0, 1, or 2), due to the relatively low failure rates. It appears that non-zero critical levels $\tilde{c}_i$, indicating that a distinction is made between customer classes, occur only in a small percentage of the SKU-s only, especially in the STEPPER case. This observation can be explained by the fact that basestock levels are low; e.g., if a basestock level is 1, a non-zero critical level means that the corresponding customer class has no access to the inventory for the item at all, i.e. zero fill rate, which is quite a rigorous differentiation measure, especially if the desired difference between the fill rates of the customer classes is small. Notice that if within a customer class all machines are identical, the expected fill rate for one machine in that class equals the obtained expected fill rate $\beta_j(c)$ of the customer class.

The most important observation for practitioners is that application of our model with customer differentiation may lead to a significant reduction in the inventory investment in comparison to the round-up policy. As can be seen from Tables 12 and 13, the investment for the STEPPER and STEP & SCAN family at ASML can be reduced with 10% and 20%, respectively, if extensive differentiation is applied, i.e. with a division into 5 customer classes. This is equal to tens of millions of EURO-s in absolute terms. Notice that in both the STEPPER and STEP & SCAN family, 80% of the cost savings of extensive differentiation can already be obtained by choosing a smart division into 2 classes. Taking into account the additional implementation efforts and the increased complexity in daily operations if 5 classes are defined, it may be wise in practice to restrict the number of customer classes.

6 Extensions

6.1 Alternative objective functions

In our model, we assumed a specific expression for inventory investment; see the objective function in the formulation of Problem (P). In this subsection, we show that our solution procedure also applies for slightly modified objective functions.

Suppose that we want to exclude pipeline stocks when assessing the inventory investment. Then the objective function in Problem (P) becomes equal to $\sum_{i=1}^{I} p_i(c_{i,J+1} - \sum_{j=1}^{J} m_{i,j}\beta_{i,j}(c_i)t_i)$, and the objective function of Problem (LR($\lambda$)) becomes equal to

$$\sum_{i=1}^{I} p_i\left(c_{i,J+1} - \sum_{j=1}^{J} m_{i,j}\beta_{i,j}(c_i)t_i\right) + \sum_{j=1}^{J} \sum_{i=1}^{I} \lambda_j m_{i,j} \left(\beta_{j,\text{obj}} - \beta_{i,j}(c_i)\right)$$

$$= \sum_{i=1}^{I} \left\{ p_i c_{i,J+1} + \sum_{j=1}^{J} (\lambda_j + t_i) m_{i,j} (1 - \beta_{i,j}(c_i)) \right\} - \sum_{i=1}^{I} \sum_{j=1}^{J} m_{i,j} t_i - \sum_{j=1}^{J} \lambda_j M_j (1 - \beta_{j,\text{obj}}).$$

For given Lagrange multipliers, the last two terms in the latter expression are constant, and the
remaining term leads to similar single-item minimization problems as before; for Problem \((SI(i))\), we obtain the same objective function as before but with the \(\lambda_j\) replaced by \(\lambda_j + t_i\). For the rest, everything remains the same in the solution procedure.

Obviously, our solution procedure also applies when the pipeline stocks have to be excluded for a subset of all SKU-s instead of for all SKU-s. Another modification that can be handled is when an extra cost occurs for each demand that is not fulfilled by the warehouse but by an emergency shipment, say (that would lead to an extra term of the form \(\sum_{i=1}^{I} a_i \sum_{j=1}^{J} m_{i,j} (1 - \beta_{i,j}(c_i))\) in the objective function of Problem \((P)\), where the \(a_i\) are item-dependent constants).

6.2 Non-identical machines with commonality

In our model, we assumed that customers have (close-to-)identical machines with (close-to-)identical failure rates; see the first and last paragraph of Section 2. We can easily relax this and apply our model to situations with different machines in different customer classes. This is relevant for real-life problems with different machine types (or families) with commonality in their material breakdown structures and a target aggregate fill rate per machine type. Then each machine type constitutes a customer class, and the whole analysis applies again. Similar results may be expected with respect to the accuracy and speed of the heuristic solution procedure. For this problem, the solution space may even be limited to basestock policies. Then it is still possible to obtain different aggregate fill rates for different machine types (if desired; the targets may also be the same for the different machine types), and one obtains a simpler solution and a simplified solution procedure.

7 Conclusions and Further Research

We have studied a multi-item, single-stage spare parts inventory model with multiple customer classes and a target aggregate fill rate per class, and critical level policies have been assumed for the inventory control. We have developed a solution procedure, based on Lagrange relaxation, for the minimization of the total investment in spare parts stocks subject to aggregate fill rate constraints. The solution procedure generates both a heuristic solution and a lower bound for the optimal costs. The procedure has been shown to be rather fast and to lead to good heuristic solutions, in particular for instances with many SKU-s. Application of the procedure to two main machine families at one representative local warehouse of ASML has shown that customer differentiation by critical level policies may lead to significant cost reductions for ASML.

The solution procedure that we developed seems appropriate for several other multi-item spare parts problems. The procedure has been successfully applied by Wong et al. [23] to a multi-item model with lateral transshipments. The procedure will be further exploited to study the use of lateral transshipments within ASML's spare parts network, and, if possible, to study multi-
item, multi-echelon models that capture lateral transshipments, emergency shipments and customer differentiation.

Acknowledgements

We would like to thank Rudi Pendavingh for proposing the method that we used for Problem \((LR)\). We thank Cor Hurkens for his suggestions regarding the AIMMS-implementation, and Ivo Adan for pointing out the paper of Dowdy et al. \([7]\) to us. Furthermore, we want to thank the company ASML, and especially Harrie de Haas, Eric Messelaar, and Harold Bol, for providing us with an interesting real-life problem as well as for their cooperation in the course of the project.

References


