Expediting in Two-Echelon Spare Parts Inventory Systems

Melvin Drent, Joachim Arts
University of Luxembourg, Luxembourg Centre for Logistics and Supply Chain Management,
melvin.drent@uni.lu, joachim.arts@uni.lu

Problem definition: We consider dual sourcing in a distribution network for spare parts consisting of one central warehouse and multiple local warehouses. Each warehouse keeps multiple types of repairable parts to maintain several types of capital goods. The repair shop at the central warehouse has two repair options for each repairable part: a regular repair option and an expedited repair option. Irrespective of the repair option, each repairable part uses a certain resource for its repair. In the design of these inventory systems, companies need to decide upon stocking levels and expedite thresholds such that total stock investments are minimized while satisfying asset availability and expediting constraints.

Academic / Practical Relevance: Although most companies have the possibility to expedite the repair of parts in short supply, no contributions have been made that incorporate such dynamic expediting policies in repairable investment decisions. Anticipating expediting decisions that will be made later leads to substantial reductions in repairable investments.

Methodology: We use queueing theory to determine the performance of the central warehouse and subsequently find the performance of all local warehouses using binomial disaggregation. For the optimization problem, we develop a greedy heuristic and a decomposition and column generation based algorithm.

Results: Both solution approaches perform very well with average optimality gaps of 2.38 and 0.27 percent, respectively, across a large test bed of industrial size. The possibility to expedite the repair of failed parts is effective in reducing stock investments with average reductions of 7.94 percent and even reductions up to 19.61 percent relative to the state of the art.

Managerial Implications: Based on a case study at Netherlands Railways, we show how managers can significantly reduce the investment in repairable spare parts when dynamic repair policies are leveraged to prioritize repair of parts whose inventory is critically low.

Key words: inventory; spare parts; multi-item; repair; expediting; multi-echelon; column generation

1. Introduction

For many industries and service organizations, the availability of capital goods such as rolling stock, manufacturing equipment and aircraft is of crucial importance for their operations. To ensure high availability of these capital goods, companies stock critical components and replace a defective component with a ready-for-use spare component after failure.

Since many critical components represent a significant financial investment, defective components are usually repaired and put back on stock rather than discarded. Consequently, the availability of capital goods largely depends on the design of the underlying spare parts inventory system for repairing and supplying these so-called repairable components.

Spare parts inventory systems for capital goods often have a two-echelon structure, in which many different types of components are stocked (Cohen et al. 1997). In this paper, we study such a multi-item two-echelon spare parts inventory system. The system consists of a set of local warehouses, i.e. operating sites, that are supported by one central warehouse. Each local warehouse maintains field inventories for spare components and sends defective components to the repair shop. The repair shop repairs these defective components and sends them to the central warehouse which replenishes the local warehouses. It is obvious that next to the inventory levels of spare components, the repair operations at the repair shop affect the availability of capital goods at the operating sites. Hence, the determination of spare components inventory levels and the design of the repair operations in the repair shop are two key aspects in the design of these two-echelon spare parts inventory systems.

In the capital goods industry, it is common practice to acquire spare components together with the acquisition of the capital good because, at that time, it is possible to negotiate reasonable prices. The determination of spare components inventory levels is therefore closely related to what is known in literature as the initial spare parts supply problem (e.g., Van Houtum and Kranenburg 2015). With respect to the repair operations at the repair shop, companies often have the flexibility to expedite the repair of defective components. Although expediting comes at an extra price, either because internal repair resources are limited or because an external repair shop charges a higher price, the possibility to expedite can significantly reduce the required initial financial investment in spare parts. Indeed, expediting the repair of defective components more often implies that a smaller initial financial investment in spare parts is required to ensure the same availability of capital goods as in spare parts inventory systems where no repair flexibility is incorporated.

Hence, in the design of the spare parts inventory systems sketched in the last two paragraphs, decision makers face two major questions:

1. How many spare parts of each repairable type should the company initially purchase and place at each warehouse?

2. When should the repair of a detective part of a given repairable type be expedited?

The objective of this paper is to present a tractable optimization model that assists decision makers in answering these two questions. These questions are faced among others by Netherlands Railways (NS), the principal Dutch passenger railway operator. Our collaboration with their maintenance department led to the present work. To establish the practical value of our optimization model, we report on a case study on their data.

As is often the case in practice, we consider a setting with several capital good types (e.g. regional trains and inter-city trains; wide-body aircraft and narrow-body aircraft) and where repairables may use different repair resources (e.g. electronic and mechanical). Because not all repairs can be expedited, many companies, including NS, use agreements between the repair shop manager and the inventory manager that determine how much of the total workload can be expedited per repair resource. Hence, the objective of our optimization model is to minimize the total investment costs in spare parts while

- not exceeding a given maximum total mean number of backorders over all local warehouses for each capital good type, and
- keeping the fraction of repairs that are expedited per repair resource below a given target level.

Because we consider critical components, a backorder for a spare part implies that the affected capital good becomes inoperable. Since failures of components typically occur very infrequently, a common assumption in the spare parts literature is that the probability that two or more backorders are from the same capital good at any point in time is negligible (e.g., Muckstadt 2005, Sherbrooke 2004). Under that assumption the average availability of a capital good type is the number of capital goods of that type minus the expected number of backorders of parts in that capital good type. As such, the first constraint of our optimization model guarantees a certain availability of each capital good type throughout the geographical region covered by the local warehouses.

In this paper, we provide a mathematical model for the decision problem described above. We assume that each local warehouse is replenished by an $(S-1,S)$ base stock policy. This means that each defective part is replaced with a ready-for-use item and is sent to the repair shop at the central warehouse immediately after the defect occurs. This replenishment policy is common in practice and is considered as well-suited for spare parts inventory control (Van Houtum and Kranenburg 2015). The central warehouse operates under an (S,T) policy similar to Song and Zipkin (2009), which keeps the usual inventory position

at constant level S, just as in a standard base stock policy. In addition, expedite threshold T triggers expedited repairs when outstanding orders in the repair pipeline are too far away. This dynamic policy thus takes into account real-time information about the repair pipeline of the repair shop, which can be obtained through modern tracking technologies. We assume that unsatisfied demand is backordered at all warehouses. Furthermore, we assume deterministic lead times for the replenishments of the local warehouses as well as for both repair options at the central warehouse.

The main contributions of this paper are summarized as follows:

1. We are the first to integrate stocking and expedited repair decisions in multi-item two-echelon spare parts inventory systems, where parts belong to different capital good types and where parts that use the same repair resource compete for expedited repair.

2. We provide a tractable optimization model that yields a tight lower bound on the optimal solution and near optimal feasible solutions. We show that our formulation allows us to decompose the non-linear non-convex integer programming problem into sub-problems per repairable type and subsequently use column generation algorithms. For the resulting sub-problem, whose state space has dimensions equal to the number of locations plus one, we provide an efficient solution algorithm that searches over only two dimensions and where each instance involves independent Newsvendor type problems.

3. As an alternative solution approach, we provide a greedy heuristic that yields excellent results. Different from most literature on greedy heuristics in spare parts inventory systems, our greedy heuristic does not only decide upon stocking levels given a certain target service level, but also on expedite thresholds such that the fraction of the total demand that receives expedited repair per repair resource remains below a certain target level.

4. Based on a case study at NS, we present insights that will help managers to understand how a dynamic repair policy can be leveraged to reduce the total investment costs in spare parts while meeting availability targets.

In his seminal paper on the METRIC model, Sherbrooke (1968) already argued that in practice, parts in short supply should be scheduled into repair first. Though, he and most contributions on the METRIC model assume that the repair lead times of each part are i.i.d. distributed, meaning that no scheduling or prioritization in repairs is possible. As a direct consequence, performance obtained in practice (either investments in stock or availability) is better than theory predicts (e.g., Rustenburg 2000, Rustenburg et al. 2001),

though exact percentages are lacking. In this paper, we are the first to relax this assumption by explicitly incorporating the possibility to change the repair lead time of a part based on the current state of the system, thereby actually scheduling parts in short supply into repair first. We show that effective usage of this possibility may lead to reductions in stock investments of up to 19.61 percent compared to static repair lead times.

The remainder of this paper is organized as follows. Section 2 reviews related literature. In Section 3, we provide a description of the model. In Section 4, we present both an exact evaluation procedure for a given control policy and the mathematical formulation of our decision problem. In Section 5, we present two solution approaches to solve this decision problem. Section 6 provides managerial insights based on a case study at NS and evaluates the performance of both solution approaches in a large test bed. Finally, some concluding remarks are presented in Section 7.

2. Literature review

Although spare parts inventory systems have been studied extensively in a variety of settings, our review involves literature with similar modeling assumptions or similar solution approaches as those used in this paper. For an extensive discussion of the existing literature in the broad field of spare parts inventory management, we refer the reader to Basten and Van Houtum (2014), Van Houtum and Kranenburg (2015) and Muckstadt (2005).

This paper contributes to the classical research line of multi-item spare parts inventory systems that started in 1968 with the seminal paper of Sherbrooke on the METRIC model. This model assumes that demand follows a Poisson process and that all warehouses operate under base stock policies. Via an approximative evaluation method, expected backorders at all local warehouses are determined for a given control policy. Since then, many extensions have been made to the METRIC model: While some researchers have focused on deriving exact steady state distributions (e.g., Graves 1985, Simon 1971), others have extended the model itself by integrating hierarchical or indentured parts structures (e.g., Muckstadt 1973), by allowing for part failures that lead to downtime after a delay (e.g., Bitton et al. 2018), or by including emergency shipments (e.g., Alfredsson and Verrijdt 1999, Lee 1987, Howard et al. 2015). The exact evaluation procedure developed in this paper shows similarities with Graves (1985). The main difference is that Graves (1985) considers only one supply mode at the central warehouse, whereas we consider both a regular and an expedited supply mode.

The system studied in this paper extends previous research which examined inventory models with multiple supply modes. We refer to Minner (2003) for an extensive discussion of such inventory models, here we discuss only the important and more relevant results. Since optimal control policies for inventory systems with expediting have complex structures (e.g., Feng et al. 2006, Whittemore and Saunders 1977), most recent papers study relatively simple heuristic policies and aim at finding (near) optimal parameters.

For single-echelon inventory systems under periodic review, an often studied heuristic policy is the dual-index policy, in which two different inventory positions are kept track off: The inventory position including arrivals within the expedited lead time and the inventory position including arrivals within the regular lead time (e.g., Arts et al. 2011, Sheopuri et al. 2010, Veeraraghavan and Scheller-Wolf 2008). Moinzadeh and Schmidt (1991) consider a similar policy for single-echelon inventory systems facing Poisson under continuous review. They focus on obtaining performance measures for a given dual-index policy when both the expedited and regular lead time are deterministic. Song and Zipkin (2009) reinterpret and extend the work of Moinzadeh and Schmidt (1991) by showing that the same inventory system with a dual-index policy and stochastic lead times is a special type of product form queueing network with one or more overflow bypasses. The dual-index policy in the setting of Moinzadeh and Schmidt (1991) and Song and Zipkin (2009) is in fact optimal for the special case where the regular repair lead time has a shifted exponential distribution and the base stock level for the regular inventory position is fixed (Arts et al. 2016). The policy that we consider for the central warehouse is equivalent to the dual-index policy of Song and Zipkin (2009). The methods of Song and Zipkin (2009) have been incorporated in a two-echelon spare parts inventory system before, albeit to decide upon emergency shipments from a so-called support warehouse to the local warehouses (Howard et al. 2015).

Literature on multiple supply modes in multi-echelon distribution systems is relatively scarce. Building upon the dual-index policy of Moinzadeh and Schmidt (1991), Moinzadeh and Aggarwal (1997) consider a two-echelon distribution system facing Poisson demand under continuous review in which all warehouses have the option to replenish their inventory through an expedited or regular supply channel. Similar to the model of Moinzadeh and Schmidt (1991), they assume deterministic lead times for both types of shipments to all warehouses. Moinzadeh and Aggarwal (1997) describe a procedure to find optimal policy

parameters and show that this system substantially improves its single-sourcing counterpart. By contrast to their paper, we consider a system where only the central warehouse has two supply modes. Yet, we impose no limitations on the lead times of those supply modes. For more variations of multi-echelon distribution systems with multiple supply modes under different cost structures and control policies, see Aggarwal and Moinzadeh (1994), Alvarez and Van der Heijden (2014), Dada (1992) and Minner et al. (2003).

Within the stream of literature focusing on inventory systems for repairable items, many contributions have been made on either expediting the repair or prioritizing the scheduling of repairs in the repair shop. As deriving structural properties of optimal policies is known to be complex when the number of different repairable types increases (Tiemessen and Van Houtum 2013), most contributions in this area resort to heuristic priority rules. We distinguish two categories of such heuristic priority rules. Under *static* priority rules, the priority of a repairable depends on its type only. Although these type of priority rules are relatively simple, several studies have shown that such rules outperform simple first come first serve rules in terms of investment costs (e.g., Adan et al. 2009, Sleptchenko et al. 2005). Under more sophisticated *dynamic* priority rules, the priority of a repairable also depends on the current state of the system. The expediting policy in our model falls into this latter category as it essentially changes the repair lead time of a part based on the current state of the repair pipeline. In a recent contribution, Arts et al. (2016) study an expediting policy similar to the present model, albeit in a single-echelon single-item setting under fluctuating demand. They remark that this expediting policy does not suffer from the tractability issues that other dynamic priority rules suffer from, while still providing the lead time flexibility inherent to this category of heuristic priority rules.

Few researchers have considered dynamic repair priority rules in multi-echelon inventory systems for repairable items. Pyke (1990) jointly addresses dynamic repair and inventory allocation decisions in a two-echelon system very similar to the one we study. He sketches a mathematical formulation of the problem to emphasize its complexity and computational intractability and subsequently resorts to simulation experiments. More recently, Caggiano et al. (2006) consider a similar problem related to dynamic repair and inventory allocation decisions. Different from the present work, their model is a finite-horizon, periodic-review model involving only one repair resource focusing on operational decisions for repairable spare parts in the exploitation phase of capital goods.

On the analysis side, we use two techniques that are widely used in the context of multi-item spare parts inventory optimization. The first technique, decomposition and column generation, is appropriate for problems that have a complicated aggregation constraint that links the different repairable types. Decomposing this problem leads to relatively simple sub-problems per repairable type. This technique has been used extensively in recent contributions on spare part inventory optimization problems (e.g., Alvarez et al. 2013, 2015, Arts 2017, Kranenburg and Van Houtum 2007, Topan et al. 2017, Wong et al. 2007). Most contributions only consider an aggregated service level as the constraint that links the different repairable types. In this paper, repairable types are not only linked through such an aggregated service level constraint, but also through the maximally allowed mean fraction of expedited repairs over all repairable types that use the same repair resource. Arts (2017) considers a similar optimization model with linking constraints on both expedited repair usage and service levels. The major difference between our work and Arts (2017) is that we consider a two-echelon spare parts inventory system. For an extensive discussion on decomposition and column generation, we refer to Dantzig and Wolfe (1960) and Lübbecke and Desrosiers (2005).

The second technique, a greedy method, is a search algorithm that iteratively selects the alternative that has the highest ratio of improvement in performance over cost increase until a feasible solution is obtained. A greedy method is quick, intuitive, easy to implement and provides satisfactory results. Although the technique has been applied in many papers on multi-item spare parts inventory optimization (e.g., Cohen et al. 1990, Kranenburg and Van Houtum 2009, Topan et al. 2017, Wong et al. 2007), none have proposed a greedy method on both stocking and expediting decisions that yields good results.

3. Model description

In this section, we first provide a brief description of the two-echelon spare parts inventory system and introduce the notation that we use throughout this paper. We then describe the policy we propose to control the system.

3.1. Description and notation

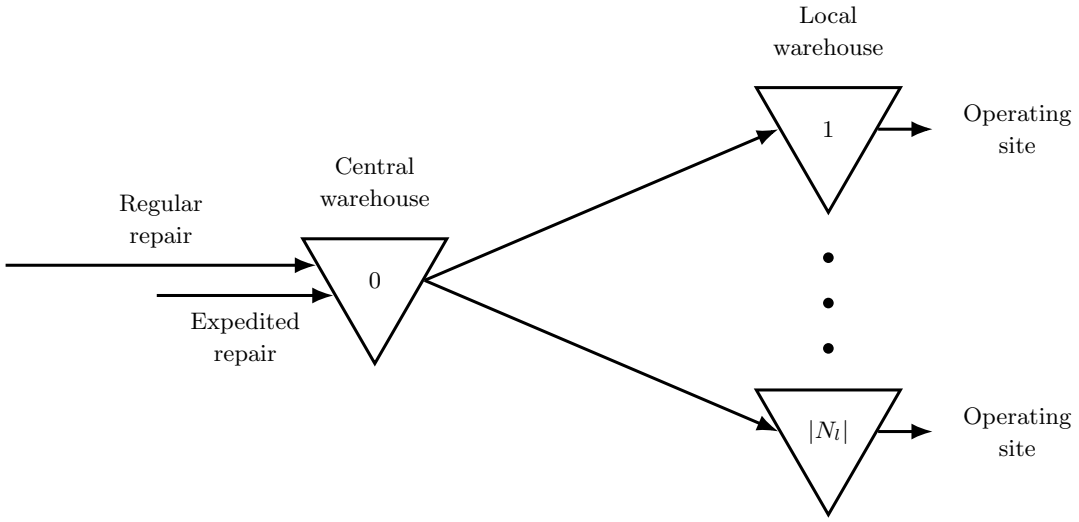
We consider a two-echelon spare parts inventory system consisting of a central warehouse and multiple local warehouses. Let the non-empty set of local warehouses be denoted by N_l. The set of all warehouses is denoted by N, i.e. $N = \{0\} \cup N_l$. Hence, the central

warehouse has index zero while the local warehouses are numbered as $n = 1, 2, \ldots, |N_l|$. Each local warehouse n is responsible for serving an operating site consisting of a number of capital goods, which may be of the same or different type. Let C denote the non-empty set of capital good types. Each capital good type $c \in C$ consists of a number of critical components that fail infrequently and independently. These critical components are crucial for operating the capital good, i.e. the capital good is down if one of these components fails. The components are at such levels in the material breakdown structure of the capital good that they can be replaced as a whole by spare parts. Component types are also called Stock Keeping Units (SKUs). Let M denote the non-empty set of critical SKUs that occur in the configurations of the different capital good types. The SKUs are numbered as $m = 1, 2, \ldots, |M|$ and each part of SKU $m \in M$ has an acquisition cost c_a^m. The set of SKUs that occur in the configuration of capital good type $c \in C$ is denoted by M_c^C. There is a set of repair resources, denoted by R, that are used to repair failed parts in the repair shop (at the central warehouse). The SKUs that use repair resource $r \in R$ in their repair are contained in the set M_r^R. We assume that M_c^C and M_r^R partition M, i.e. $\cup_{c \in C} M_c^C = \cup_{r \in R} M_r^R = M$ and $\cap_{c \in C} M_c^C = \cap_{r \in R} M_r^R = \emptyset$. This assumption is common in practice and simplifies notation considerably; it is however not essential to our analysis. As there might be settings where this assumption does not hold, Online Appendix D briefly shows how this assumption can be readily relaxed along similar lines as is done in Kranenburg and Van Houtum (2007).

Demand for SKU m at local warehouse n follows is a Poisson process with rate $\lambda_{m,n}$. This demand model is common in literature and accurate in practice for spare parts (Caglar et al. 2004, Sherbrooke 1968, Graves 1985). When a demand for SKU $m \in M$ occurs at local warehouse $n \in N_l$, it will be filled from stock, or backordered if the stock is depleted. In the latter case, the capital good remains down until a spare part becomes available at the local warehouse. The failed part is shipped to the repair shop at the central warehouse, where all failed parts are immediately sent into regular repair or expedited repair, where the corresponding repair resource $r \in R$ is used for repair. At the same time, the central warehouse ships a spare part to the local warehouse from its inventory, if it has an available spare part. Otherwise, the replenishment order is backordered at the central warehouse until a part is repaired and becomes available. Upon completion of repair a part is put back on stock at the central warehouse.

The order and shipment time for a spare part of SKU m from the central warehouse to local warehouse n is fixed and denoted by $t_{m,n}$. Note that $t_{m,n}$ excludes any waiting time at the central depot when a spare part is not available. For returned failed parts at the repair shop, it takes either $t^{reg}_{m,0}$ time units, in case of the regular repair, or $t^{exp}_{m,0}$ time units, in case of the expedited repair, until the part is returned to the spare parts stock at the central warehouse. We assume that both repair times are fixed, with $t^{reg}_{m,0} > t^{exp}_{m,0} > 0$. Figure 1 provides a graphical representation of the system under consideration and notation is summarized in Table 1 (including notation introduced later).

Figure 1 Two-echelon spare parts inventory system with expediting



3.2. Control policy

Each failed part at a local warehouse results in an immediate replenishment order at the central warehouse. This implies that the inventory positions of a given SKU $m \in M$ remain constant at all local warehouses. Hence, we have base stock control at each local warehouse $n \in N_l$ for each SKU $m \in M$ and we denote the corresponding base stock levels by $S_{m,n}$.

The central warehous is controlled by a dual-index policy. This policy has two parameters for each SKU $m \in M$, integers $S_{m,0}$ and $S'_{m,0}$, with $S_{m,0} \geq S'_{m,0}$. Let $t^1_{m,0} = t^{reg}_{m,0} - t^{exp}_{m,0}$, i.e. the additional regular lead time, and $t^2_{m,0} = t^{exp}_{m,0}$. We define two inventory positions for each SKU m: $IP^1_{m,0}$ and $IP^2_{m,0}$. $IP^1_{m,0}$ is the usual local inventory position and includes net inventory $IN_{m,0}$ (on-hand stock $OH_{m,0}$ minus any backorders $BO_{m,0}$) plus all parts in repair $X_{m,0}$. $IP^2_{m,0}$ is similar but only includes those parts in repair $X^2_{m,0}$ that will be

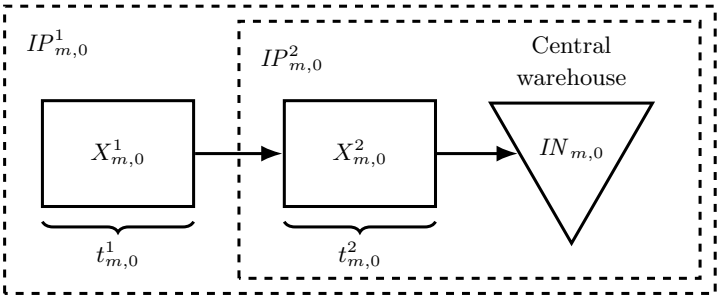
Table 1 Overview of notation

Notation	Description
Sets	
N	Set of all warehouses.
$N_l \subset N$	Set of local warehouses.
M	Set of all SKUs.
C	Set of capital good types.
R	Set of repair resources.
$M_r^R \subseteq M$	Set of SKUs that use repair resource $r \in R$ in the repair of failed parts.
$M_c^C \subseteq M$	Set of SKUs that occur in the configuration of capital good type $c \in C$.
Input parameters	
$\lambda_{m,n}$	Demand intensity for SKU $m \in M$ at warehouse $n \in N$.
δ_m^r	Fraction of demands over all parts of SKUs $k \in M_r^R$ that are from SKU m, i.e. $\frac{\lambda_{m,0}}{\sum_{k \in M_r^R} \lambda_{k,0}}$.
$t_{m,n}$	Lead time from the central warehouse to local warehouse $n \in N_l$ of SKU $m \in M$.
$t_{m,0}^{reg}$	Regular repair lead time of SKU $m \in M$.
$t_{m,0}^{exp}$	Expedited repair lead time of SKU $m \in M$, also denoted by $t_{m,0}^2$.
$t_{m,0}^1$	Additional regular repair lead time of SKU $m \in M$.
c_a^m	Acquisition cost for SKU $m \in M$.
$\mathcal{B}_c^{max}$	The maximally allowed mean number of backorders over all SKUs $m \in M_c^C$ for capital good type $c \in C$.
$\mathcal{E}_r^{max}$	The maximally allowed mean fraction of expedited repairs over all SKUs $m \in M_r^R$ that use repair resource $r \in R$ during their repair.
Decision variables	
$S_{m,n}$	Base stock level of SKU $m \in M$ at warehouse $n \in N$.
$\mathbf{S}_m$	The vector $(S_{m,0}, S_{m,1}, \ldots, S_{m,\lvert N_l \rvert})$.
$\mathbf{S}$	The base stock levels matrix $[S_{m,n}]$.
T_m	Expedite threshold of SKU $m \in M$.
T	The vector $(T_1, T_2, \ldots, T_{\lvert M \rvert})$.
State variables	
$X_{m,0}$	Number of outstanding repairs of SKU $m \in M$ at the central warehouse.
$X_{m,0}^2$	Number of outstanding repairs of SKU $m \in M$ at the central warehouse that will be repaired within $t_{m,0}^2$ time units.
$X_{m,0}^1$	Number of outstanding repairs of SKU $m \in M$ at the central warehouse that will not be repaired within $t_{m,0}^2$ time units, i.e. $X_{m,0}$-$X_{m,0}^2$.
$X_{m,n}$	Number of outstanding orders of SKU $m \in M$ at local warehouse $n \in N_l$.
Output of model	
$EBO_{m,n}(\mathbf{S}_m, T_m)$	Mean number of backorders for SKU m at local warehouse $n \in N_l$ under a given control policy $(\mathbf{S}_m, T_m)$, i.e. $\sum_{x=S_{m,n}+1}^{\infty}(x - S_{m,n})\mathbb{P}\{X_{m,n} = x\}$.
$EBO_c(\mathbf{S}, T)$	Aggregate mean number of backorders for capital good type $c \in C$ under a given control policy $(\mathbf{S}, T)$, i.e. $\sum_{m \in M_c^C} \sum_{n \in N_l} EBO_{m,n}(\mathbf{S}_m, T_m)$.
$EXP_m(T_m)$	Fraction of failed parts of SKU $m \in M$ that utilize the expedited repair option under a given expedite threshold (T_m), i.e. $\mathbb{P}\{X_{m,0}^1 = T_m\}$.
$EXP_r(T)$	Aggregate mean fraction of failed parts over all SKUs $m \in M_r^R$ that utilize the expedited repair option under a given expedite threshold vector (T), i.e. $\sum_{m \in M_r^R} \delta_m^r EXP_m(T_m)$.
$C(\mathbf{S})$	The total investment costs in spare parts under a given base stock levels matrix $\mathbf{S}$, i.e. $\sum_{m \in M} \sum_{n \in N} c_a^m S_{m,n}$.
C_P^{UB} (C_P^{LB})	Upper (lower) bound for the optimal solution to problem (P) in (3)-(6).
C_{BENCH}^{LB}	Lower bound for the optimal solution of a benchmark instance.

repaired and returned to on-hand stock within $t_{m,0}^2$ time units. Hence, the number of parts in repair that will not be repaired and returned to on-hand stock within $t_{m,0}^2$ time units $X_{m,0}^1$ is equal to $X_{m,0} - X_{m,0}^2$. Figure 2 provides a graphical representation of the two different inventory positions at the central warehouse.

The dual-index policy works as follows: Keep $IP_{m,0}^1$ at constant level $S_{m,0}$ (as in standard base stock control) and also $IP_{m,0}^2 \geq S'_{m,0}$. Thus upon the demand of a part and the return of a failed part of SKU $m \in M$, we first examine $IP_{m,0}^2$. If $IP_{m,0}^2$ (after the failed part is returned, but before deciding upon the repair option) is already $S'_{m,0}$ or greater, we send it

Figure 2 **Inventory positions at the central warehouse**



into regular repair. However, if a regular repair would leave $IP^2_{m,0} < S'_{m,0}$, then we use the expedited repair option. Note that $IP^2_{m,0} = IN_{m,0} + X^2_{m,0} = IP^1_{m,0} - X^1_{m,0} = S_{m,0} - X^1_{m,0}$, and thus, equivalently, the dual-index policy keeps $S_{m,0} - X^1_{m,0} \geq S'_{m,0}$. Hence, defining expedite threshold $T_m = S_{m,0} - S'_{m,0}$ $\forall m \in M$, the dual-index policy sends failed parts into regular repair as long as $X^1_{m,0} \leq T_m$ (cf. Song and Zipkin 2009).

Let $\mathbf{S}_m = (S_{m,0}, S_{m,1}, \ldots, S_{m,|N_l|})$, $m \in M$, denote the vector of base stock levels for SKU m. Then, a control policy $(\mathbf{S}, T)$ is denoted by base stock levels matrix $\mathbf{S}$ and a vector $T = (T_1, T_2, \ldots, T_{|M|})$ containing the expedite thresholds of each SKU $m \in M$.

4. Performance evaluation and problem formulation

In this section, we first provide an exact evaluation procedure for a given control policy. Subsequently, we present the mathematical formulation of our decision problem.

4.1. Exact evaluation of a given control policy

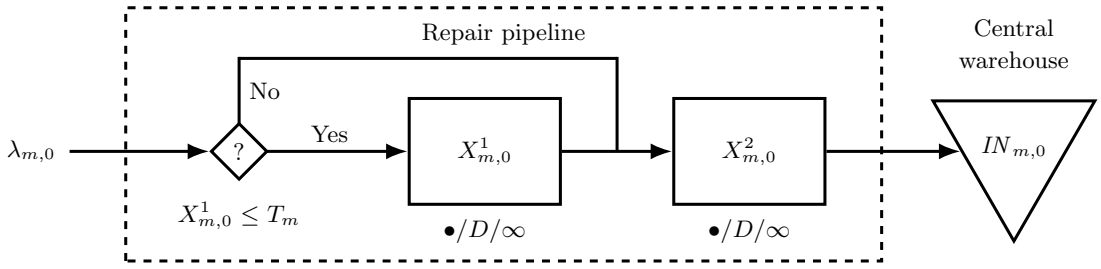
The evaluation of a given control policy $(\mathbf{S}, T)$ can be done per SKU. Consider therefore some SKU $m \in M$ that has base stock vector $\mathbf{S}_m$ and expedite threshold T_m. We first consider the performance of SKU m at the central warehouse, and subsequently link this to its performance at all local warehouses.

Key in evaluating the performance of the central warehouse for SKU m is to obtain the distribution of the number of parts in repair $X_{m,0}$. Since each failure of SKU m results in an immediate replenishment request for SKU m at the central warehouse, the demand process for parts of SKU m as seen by the central warehouse is a Poisson process with constant rate $\lambda_{m,0} = \sum_{n \in N_l} \lambda_{m,n}$. Each replenishment request for SKU m is accompanied by a failed part that goes into repair. Hence, failed parts of SKU m enter the repair pipeline according to a Poisson process with constant rate $\lambda_{m,0}$. The fraction of demands for SKU

m over demands from all SKUs that use the same repair resource $r \in R$ as SKU m uses, is then given by $\delta_m^r = \frac{\lambda_{m,0}}{\sum_{k \in M_r^R} \lambda_{k,0}}$.

Under the dual-index policy described in the former section, the repair pipeline of each SKU m can then be seen as an open queueing network with outside Poisson arrivals at constant rate $\lambda_{m,0}$, and two $\bullet/D/\infty$ queues that produce delays of t_m^1 and t_m^2 time units. Figure 3 provides a graphical representation of this open queueing network.

Figure 3 Repair pipeline as an open queueing network



Observe that a normal repair first passes through queue 1, where it remains t_m^1 units, and then passes through queue 2, where it remains t_m^2 units. Failed parts that receive expedited repair bypass the first part of the repair pipeline and only pass through queue 2. After a part completes queue 2, it arrives to inventory at the central warehouse. In effect, the dual-index policy directs failed parts of SKU m into normal repair as long as the number of repairs at queue 1, i.e. $X_{m,0}^1$, is not greater than the expedite threshold T_m. When an arriving failed part would overflow T_m, the failed part bypasses queue 1 and goes directly to queue 2, that is, the failed part goes into expedited repair. In the queueing literature, this is sometimes referred to as *jump over blocking* or as an *overflow bypass* (e.g., Lam 1977, Song and Zipkin 2009).

The distribution of the number of parts in repair $X_{m,0}$, follows from the joint distribution of $(X_{m,0}^1, X_{m,0}^2)$. Let $p_m(x_1, x_2) = \mathbb{P}\{X_{m,0}^1 = x_1, X_{m,0}^2 = x_2\}$ denote the steady-state joint distribution of $(X_{m,0}^1, X_{m,0}^2)$. Let $\phi_m^1(x_1)$ and $\phi_m^2(x_2)$ denote the Poisson probabilities $e^{\lambda_{m,0} \cdot t_m^1}(\lambda_{m,0} \cdot t_m^1)^{x_1}/x_1!$ and $e^{\lambda_{m,0} \cdot t_m^2}(\lambda_{m,0} \cdot t_m^2)^{x_2}/x_2!$, respectively. The support of $(X_{m,0}^1, X_{m,0}^2)$ is denoted by $\mathscr{X}(T_m) = \{(x_1, x_2) \in \mathbb{N}_0^2 : x_1 \leq T_m\}$. Then, as shown by Lam (1977) and Song and Zipkin (2009), the joint distribution of $(X_{m,0}^1, X_{m,0}^2)$ has product-form

$$p_m(x_1, x_2) = \frac{\phi_m^1(x_1)\phi_m^2(x_2)}{\sum_{x_1 \leq T_m} \phi_m^1(x_1)}, \qquad (x_1, x_2) \in \mathscr{X}(T_m).$$

Letting $p_m(x) = \mathbb{P}\{X_{m,0} = x\}$ denote the equilibrium probability of the number of parts in repair $X_{m,0}$ and $\phi_m(x)$ denote the Poisson probability $e^{\lambda_{m,0} \cdot t_m^{reg}} (\lambda_{m,0} \cdot t_m^{reg})^x / x!$, we obtain

$$p_m(x) = \sum_{i=0}^{x} p_m(i, x-i) = \left(\sum_{x_1 \leq T_m} \phi_m^1(x_1) \right)^{-1} \begin{cases} \phi_m(x) & x \leq T_m, \\ \sum_{i=0}^{T_m} \phi_m^1(i) \phi_m^2(x-i) & x > T_m. \end{cases}$$

The main performance measures of the central warehouse are now easily obtained. For SKU m, the on-hand stock $OH_{m,0}$ and the number of backorders $BO_{m,0}$ are equal to $(S_{m,0} - X_{m,0})^+$ and $(X_{m,0} - S_{m,0})^+$, respectively, where $x^+ = \max(0, x)$. For their probability distributions, we obtain

$$\mathbb{P}\{OH_{m,0} = x\} = \begin{cases} \sum_{j=S_{m,0}}^{\infty} p_m(j), & x = 0, \\ p_m(S_{m,0} - x), & x \in \{1, \ldots, S_{m,0}\}, \end{cases}$$

$$\mathbb{P}\{BO_{m,0} = x\} = \begin{cases} \sum_{j=0}^{S_{m,0}} p_m(j), & x = 0, \\ p_m(S_{m,0} + x), & x > 0. \end{cases}$$

In addition, letting $\rho_m = \lambda_{m,0} \cdot t_{m,0}^1$, the fraction of failed parts of SKU m that utilize the expedite repair option is given by

$$EXP_m(T_m) = \mathbb{P}\{X_{m,0}^1 = T_m\} = \frac{\rho_m^{T_m}}{T_m!} \left(\sum_{i=0}^{T_m} \frac{\rho_m^i}{i!} \right)^{-1},$$

which is the (Erlang) blocking probability of an $M/G/c/c$ queue, where the numbers of parallel servers c is equal to expedite threshold T_m (e.g., Gross et al. 2008).

Key in evaluating the performance of each local warehouse $n \in N_l$ for SKU m is to obtain the distribution of orders outstanding for each local warehouse. Therefore we need to determine the distribution of backorders for a SKU m at the central warehouse that belong to local warehouse n.

Simon (1971) shows that when outstanding orders at the central warehouse are filled on a first-come first-served bases, then each backorder at the central warehouse belongs to local warehouse n with probability $\frac{\lambda_{m,n}}{\lambda_{m,0}}$, independently across backorders. Let $BO_{m,0}^n$ denote the number of backorders of local warehouse n in the backorder queue of SKU m at the central warehouse. Then, by conditioning on the number of backorders of SKU m at

the central warehouse and using Simon's result that the conditional distribution of $BO^n_{m,0}$ is a binomial distribution, we obtain the following probability distribution for this number of backorders

$$\begin{aligned}\mathbb{P}\{BO^n_{m,0}=x\} &= \sum_{y=x}^{\infty}\mathbb{P}\{BO^n_{m,0}=x|BO_{m,0}=y\}\mathbb{P}\{BO_{m,0}=y\}\\ &= \sum_{y=x}^{\infty}\binom{y}{x}\left(\frac{\lambda_{m,n}}{\lambda_{m,0}}\right)^x\left(1-\frac{\lambda_{m,n}}{\lambda_{m,0}}\right)^{y-x}\mathbb{P}\{BO_{m,0}=y\}. \qquad (1)\end{aligned}$$

Now, we determine the distribution of the outstanding orders of SKU m at each local warehouse n. The outstanding orders at any time t consists of demand that occurred in the interval $(t-t_{m,n},t)$ (notation $D_{m,n}(t-t_{m,n},t)$) and backorders at the central warehouse that belong to local warehouse n at time $t-t_{m,n}$ (notation $BO^n_{m,0}(t)$), i.e. $X_{m,n}(t)=D_{m,n}(t-t_{m,n},t)+BO^n_{m,0}(t)$. Since the Poisson process has independent increments, $D_{m,n}(t-t_{m,n},t)$ and $BO^n_{m,0}(t)$ are independent random variables so that in stationary state

$$X_{m,n}=D_{m,n}+BO^n_{m,0},$$

where $D_{m,n}$ has a Poisson distribution with mean $\lambda_{m,n}t_{m,n}$ and the distribution of $BO^n_{m,0}$ is given in (1). Therefore the stationary distribution of $X_{m,n}$ is obtained by convolution.

From this point, the main performance measures of each local warehouse are easily obtained. For SKU m, the on-hand stock $OH_{m,n}$ and the number of backorders $BO_{m,n}$ at local warehouse n are equal to $(S_{m,n}-X_{m,n})^+$ and $(X_{m,n}-S_{m,n})^+$, respectively. Their probability distributions are then obtained in a similar way as the probability distributions of the on-hand stock and the number of backorders at the central warehouse. In particular, the mean number of backorders for SKU m at local warehouse n is given by

$$EBO_{m,n}(\mathbf{S}_m,T_m)=\sum_{x=S_{m,n}+1}^{\infty}(x-S_{m,n})\mathbb{P}\{X_{m,n}=x\}. \qquad (2)$$

4.2. Problem formulation

For a given control policy $(\mathbf{S},T)$, we denote the total investment costs in spare parts as

$$C(\mathbf{S})=\sum_{m\in M}\sum_{n\in N}c^m_a S_{m,n},$$

the aggregate mean number of backorders for capital good type c as

$$EBO_c(\mathbf{S},T)=\sum_{m\in M^C_c}\sum_{n\in N_l}EBO_{m,n}(\mathbf{S}_m,T_m),$$

and the aggregate mean fraction of failed parts of all SKUs $m \in M_r^R$ that utilize the expedited repair option using repair resource $r \in R$ as

$$EXP_r(T) = \sum_{m \in M_r^R} \delta_m^r EXP_m(T_m).$$

The objective of our decision problem is to minimize the total investment costs in spare parts while keeping the mean number of aggregate backorders for each capital good type $c \in C$ below $\mathcal{B}_c^{max}$ and keeping the fraction of repairs that are expedited per repair resource $r \in R$ below $\mathcal{E}_r^{max}$. Combining the aforementioned results in the following mathematical formulation of our decision problem which we call problem (P):

$$(P) \qquad \min_{\{\mathbf{S},T\}} \quad C(\mathbf{S}) \tag{3}$$

$$\text{subject to} \quad EBO_c(\mathbf{S},T) \leq \mathcal{B}_c^{max}, \qquad \forall c \in C \tag{4}$$

$$EXP_r(T) \leq \mathcal{E}_r^{max}, \qquad \forall r \in R \tag{5}$$

$$\mathbf{S} \in \mathscr{S}, \quad T \in \mathbb{N}_0^{|M|}, \tag{6}$$

where $\mathscr{S} = \{\mathbf{S} : S_{m,n} \in \mathbb{N}_0,\ \forall m \in M \text{ and } \forall n \in N\}$. Let $(\mathbf{S}^*, T^*)$ denote an optimal solution to problem (P) and let C_P be the corresponding optimal cost.

We conclude this section with two remarks related to problem (P). First, note that problem (P) can be considered as a non-linear non-convex knapsack problem with multiple constraints, where more than one copy of each item can be selected. It is well-known that even the simplest type of knapsack problems belongs to the class of $\mathcal{NP}$-hard problems (Kellerer et al. 2004). As our knapsack problem is more complex, it is very likely that also for problem (P) no polynomial time optimization algorithm exists.

Second, we emphasize that constraint (5) models agreements between repair shop managers and inventory managers that determine how much of the total stream of failed parts can be expedited per repair resource. This constraint therefore has practical and intuitive appeal. There might be settings where it is relatively easy to obtain the exact costs associated with expediting a repair (e.g. in case of an external repair shop). In Online Appendix C, we show how the subsequent analysis is readily extended to the setting where additional costs are charged for expediting repairs.

5. Optimization of base stock levels and expedite thresholds

In this section, we provide two solution approaches for problem (P). We first present a decomposition and column generation (DCG) algorithm to construct a lower bound for problem (P). We then show how the sub-problem of this algorithm can be solved efficiently. Subsequently, we show how to find a good feasible solution for problem (P). Finally, we describe a greedy algorithm for problem (P).

5.1. Constructing lower bounds

We first reformulate problem (P) as a partitioning problem so that we can apply the technique of column generation (also known as Dantzig-Wolfe decompostion). This technique was pioneered by Dantzig and Wolfe (1960) and a thorough modern treatment is given by Lübbecke and Desrosiers (2005). Thus we obtain an integer linear program for which we relax the integrality constraints. We refer to this problem as the master problem (MP). Let K_m be the set of all policies k for SKU $m \in M$ that respect constraint (6) of problem (P). Each policy $k \in K_m$ has base stock vector $\mathbf{S}_m^k := (S_{m,0}^k, S_{m,1}^k, \ldots, S_{m,|N_l|}^k)$ and expedite threshold T_m^k. Let $x_m^k \in \{0,1\}$, $m \in M$, $k \in K_m$, denote the decision variable indicating whether policy k is chosen ($x_m^k = 1$) for SKU m or not ($x_m^k = 0$). Then, by relaxing the integrality constraint on x_m^k, the master problem (MP) is defined as follows:

$$(MP) \quad \min_{\{x_m^k : m \in M, k \in K_m\}} \sum_{m \in M} \sum_{n \in N} \sum_{k \in K_m} c_a^m S_{m,n}^k x_m^k \tag{7}$$

$$\text{subject to} \quad \sum_{m \in M_c^C} \sum_{n \in N_l} \sum_{k \in K_m} EBO_{m,n}(\mathbf{S}_m^k, T_m^k) x_m^k \leq \mathcal{B}_c^{max}, \quad \forall c \in C \tag{8}$$

$$\sum_{m \in M_r^R} \sum_{k \in K_m} \delta_m^r EXP_m(T_m^k) x_m^k \leq \mathcal{E}_r^{max}, \quad \forall r \in R \tag{9}$$

$$\sum_{k \in K_m} x_m^k = 1, \quad \forall m \in M \tag{10}$$

$$x_m^k \geq 0, \quad \forall m \in M, \forall k \in K_m \tag{11}$$

Let C_P^{LB} denote the optimal cost for master problem (MP). Due to the relaxation of the integrality constraint on x_m^k, an optimal cost C_P^{LB} is also a lower bound on the optimal cost for problem (P), C_P.

Since the set K_m contains an infinite number of policies, a restricted master problem (RMP) is introduced in which, for each SKU $m \in M$, only a small subset of policies $K_m^{res} \subseteq K_m$ is considered. After solving (RMP) to optimality, we are interested in policies $K_m \setminus K_m^{res}$

that will improve the solution of (RMP) if they are added. To check whether such policies exist, we solve, for each SKU m, a column generation sub-problem. To this end, let p_c denote the dual variable of (RMP) corresponding with the expected backorder constraint (8) for capital good type $c \in C$, let ρ_r denote the dual variable of (RMP) corresponding with the expected fraction of expedited repairs constraint (9) for repair resource $r \in R$ and let υ_m denote the dual variable of (RMP) corresponding to constraint (10) that assures that for each SKU $m \in M$ a convex combination of policies is chosen. Then, the column generation sub-problem for SKU $m \in M_r^R \cap M_c^C$ of (RMP) is given by:

$$(SUB(m)) \quad \min_{\{(\mathbf{S}_m, T_m)\}} \quad \sum_{n \in N} c_a^m S_{m,n} - p_c \sum_{n \in N_l} EBO_{m,n}(\mathbf{S}_m, T_m) - \rho_r \delta_m^r EXP_m(T_m) - \upsilon_m \tag{12}$$

$$\text{subject to} \qquad \mathbf{S}_m \in \mathbb{N}_0^{|N|}, \qquad T_m \in n \in \mathbb{N}_0. \tag{13}$$

If a feasible solution to $(SUB(m))$ exists with a negative objective value, then the objective of (RMP) can be improved by adding this policy to K_m^{res} and solving (RMP) with the larger set K_m^{res}. An optimal solution for (RMP) is also an optimal solution for (MP) if for none of the SKUs a policy with negative reduced costs exists.

In the next section, we present an exact solution method to solve $(SUB(m))$. However, we remark that all policies that yield a negative objective value for $(SUB(m))$, can improve the solution of (RMP). Hence, we do not necessarily have to solve $(SUB(m))$ to optimality each time we obtain new dual variables from (RMP).

5.2. Solving the sub-problem

This section treats an exact solution method for $(SUB(m))$. All proofs are in Online Appendix A. If we fix the control policy parameters at the central warehouse, then this warehouse simply becomes a supplier with a known stochastic lead time from the perspective of each local warehouse. Hence, for fixed T_m and $S_{m,0}$, each local warehouse $n \in N_l$ operates as an independent Newsvendor subsystem, and we can optimize them separately:

THEOREM 1. *The optimal $S_{m,n}$, $n \in N_l$, for fixed values of $S_{m,0}$ and T_m, $S_{m,n}^*(S_{m,0}, T_m)$, is the smallest $S_{m,n}(S_{m,0}, T_m)$ that satisfies*

$$P\{X_{m,n}(S_{m,0}, T_m) \leq S_{m,n}(S_{m,0}, T_m)\} \geq \frac{p_c + c_a^m}{p_c}. \tag{14}$$

The remaining problem of finding the optimal control policy parameters at the central warehouse is more involved. In fact, it is known that objective function (12) is not convex in $S_{m,0}$ for a fixed T_m and corresponding $S^*_{m,n}$, $n \in N_l$ (e.g., Gallego et al. 2007, Rong et al. 2017). Similarly, it can readily be verified that the objective function (12) is also not convex in T_m for a fixed $S_{m,0}$ and corresponding $S^*_{m,n}$, $n \in N_l$. Finding the optimal control policy parameters at the central warehouse therefore requires an enumerative search.

To simplify this search, we now proceed with establishing an upper bound on the optimal base stock level at the central warehouse for a given expedite threshold. If the expedite threshold is fixed and the local warehouses carry no inventories, then only the base stock level at the central warehouse can influence the expected backorders at all local warehouses. Hence, the following lemma shows that for fixed T_m and $S_{m,n} = 0 \ \forall n \in N_l$, the central warehouse also operates as an independent Newsvendor subsystem:

LEMMA 1. *The optimal $S_{m,0}$ for fixed T_m and $S_{m,n} = 0$ for all $n \in N_l$, say $\bar{S}_{m,0}(T_m)$, is the smallest $S_{m,0}$ that satisfies*

$$P\{X_{m,0}(T_m) \leq S_{m,0}\} \geq \frac{p_c + c_a^m}{p_c}. \tag{15}$$

Observe that if the local warehouses increase their base stock levels, then the amount of inventory that the central warehouse should carry can only decrease (assuming that the expedite threshold is fixed). It is therefore clear that $\bar{S}_{m,0}(T_m)$ obtained using Lemma 1 is in fact an upper bound on $S^*_{m,0}(T_m)$ because it assumes no inventories at the local warehouses. This is formalized in the next two results.

LEMMA 2. *Let $S^*_{m,n}(S_{m,0}, T_m)$ be the optimal value of $S_{m,n}$, $n \in N_l$, for given values of T_m and $S_{m,0}$. Then $S^*_{m,n}(S_{m,0}, T_m)$ is non-increasing in $S_{m,0}$.*

THEOREM 2. *$\bar{S}_{m,0}(T_m)$, as specified in Lemma 1, is an upper bound for $S^*_{m,0}(T_m)$.*

Based on the results presented above, we propose the following exact solution method to solve $(SUB(m))$. We set T_m to 0 and then search over T_m. For each value of T_m, we vary $S_{m,0}$ over $0 \leq S_{m,0} \leq \bar{S}_{m,0}(T_m)$, where $\bar{S}_{m,0}(T_m)$ is determined using Lemma 1. For each pair $(S_{m,0}, T_m)$, we optimize $S_{m,n}$ for all $n \in N_l$ using Theorem 1. Since the objective function of $(SUB(m))$ for fixed values of $S_{m,0}$ and corresponding $S^*_{m,n}(S_{m,0}, T_m)$, $n \in N_l$, is not convex in T_m, we continue the search over T_m by examining a few values beyond the last observed local minimum.

5.3. Constructing a good feasible solution

When no more policies can be added to K_m^{res}, then a solution to the final version of problem (RMP) provides a lower bound, C_P^{LB}, on the optimal cost for problem (P), C_P. In case there are no fractional solutions for any x_m^k, $m \in M$, $k \in K_m$, this also is an upper bound, C_P^{UB}, for C_P. If there are fractional solutions for any x_m^k, we solve the final version of problem (RMP) as an integer linear program. Alvarez et al. (2013, 2015) show that this approach yields very good results compared to other methods such as local search algorithms. To speed up the solution process of solving the final version of problem (RMP) as an integer linear program, we use the feasibility pump heuristic of Fischetti et al. (2005), and we stop the solution of the integer linear program as soon as a feasible solution with optimality gap of less than 0.5 percent is found or 1 minute has elapsed (whichever occurs first). This results in a good feasible solution to problem (P). The corresponding cost of this solution is also an upper bound, C_P^{UB}, for C_P.

Pseudo-code of the DCG algorithm as well as the greedy heuristic described in the next section can be found in Online Appendix B.

5.4. A two-step greedy approach

We now describe a greedy heuristic for problem (P). This greedy heuristic consists of two steps that are executed consecutively. In the first step, we determine, independent of base stock level matrix $\mathbf{S}$, expedite threshold vector T. Subsequently, based on the vector of expedite thresholds T determined in the first step, we find base stock levels matrix $\mathbf{S}$.

Expediting the repair of a given SKU $m \in M$ implies that fewer parts of m are needed to provide the same availability as when no repairs are expedited. Hence, given that repair resources are limited, we want to expedite the repair of expensive parts more often than cheaper parts. In addition, the cost benefit of expediting the repair of a given SKU $m \in M$ increases in its additional regular lead time, i.e. $t_{m,0}^1$. Hence, given that repair resources are limited, we want to expedite the repair of parts with a greater additional regular repair lead time more often than parts with a smaller additional regular repair lead time.

If there were no restrictions on the aggregate mean fractions of failed parts that are expedited, then, irrespective of base stock levels matrix $\mathbf{S}$, the zero vector would be the optimal vector of expedite thresholds. Hence, in the first step of the greedy heuristic, we set all expedite thresholds T_m, $m \in M$, to zero and then start with greedy steps, in which

we increase T_m leading to the largest decrease in distance to the set of feasible expedite vectors per acquisition cost and additional regular repair lead time.

The first step of the greedy heuristic is formally described as follows. We first partition the set of all expedite thresholds vectors T into a subset T^{feas} of feasible expedite thresholds vectors, i.e. that respect constraint (5) of problem (P), and a subset $\mathbb{N}_0^{|M|} \setminus T^{feas}$ of infeasible expedite thresholds vectors. Next, for each expedite thresholds vector, we define the distance $d(T)$ to T^{feas} as

$$d(T) = \sum_{r \in R} (EXP_r(T) - \mathcal{E}_r^{max})^+ .$$

In each greedy step, we have a current solution $T \in \mathbb{N}_0^{|M|} \setminus T^{feas}$, and we look at the ratio of the decrease in distance to T^{feas} if T_m, $m \in M$, is increased by one unit and the product of the acquisition cost and the additional regular repair lead time.

Let $-\Delta_m d(T)$ denote the decrease in distance to the set of feasible vectors of expedite thresholds. For a given SKU $m \in M$ that uses repair resource $r \in R$ in the repair of its failed parts, we obtain

$$\begin{aligned} \Delta_m d(T) &= d(T + \mathbf{e}_m) - d(T), \\ &= (EXP_r(T + \mathbf{e}_m) - \mathcal{E}_r^{max})^+ - (EXP_r(T) - \mathcal{E}_r^{max})^+ , \end{aligned}$$

where $\mathbf{e}_m$ is a $|M|$-dimensional vector with a 1 on position m (the positions are numbered as $0, 1, \ldots, |M|-1$) and zero on the other positions.

Since the Erlang loss formula, and thus $EXP(T_m)$, is convex and decreasing in T_m (e.g., Messerli 1972), it follows that $-\Delta_m d(T) \geq 0$ for all $m \in M$. The ratio $\Gamma_m^T = \frac{-\Delta_m d(T)}{t_{m,0}^1 c_a^m}$ denotes the decrease in distance to the set of feasible vectors of expedite thresholds per both the acquisition cost and the additional regular repair lead time. During each greedy step, we increase the expedite threshold of SKU m with the highest Γ_m^T to $T_m + 1$. We continue with these steps until we arrive at a feasible solution T and we denote this solution by $\bar{T}$.

The second step of the greedy heuristic is motivated by Wong et al. (2007). If there were no restrictions on the aggregate mean numbers of backorders, then, irrespective of the vector of expedite thresholds, the zero matrix would be the optimal base stock levels matrix. Hence, in the second step of the greedy heuristic, we set all base stock levels $S_{m,n}$, $m \in M$, $n \in N$, to zero and then start with greedy steps, in which we increase $S_{m,n}$ leading

to the largest decrease in distance to the set of feasible base stock levels matrices per acquisition cost.

The second step of the greedy heuristic is formally described as follows. We first partition the set of all base stock levels matrices $\mathscr{S}$ into a subset $\mathscr{S}^{feas}$ of feasible base stock levels matrices, i.e. that respect constraint (4) of problem (P), and a subset $\mathscr{S} \setminus \mathscr{S}^{feas}$ of infeasible base stock levels matrices. Next, for each base stock levels matrix, we define the distance $d(\mathbf{S}, \bar{T})$ to $\mathscr{S}^{feas}$ as

$$d(\mathbf{S}, \bar{T}) = \sum_{c \in C} \left(EBO_c(\mathbf{S}, \bar{T}) - \mathcal{B}_c^{max} \right)^+ .$$

In each greedy step, we have a current solution $\mathbf{S} \in \mathscr{S} \setminus \mathscr{S}^{feas}$, and we look at the ratio of the decrease in distance to $\mathscr{S}^{feas}$ and the acquisition cost if $S_{m,n}$, $m \in M$, $n \in N$, is increased by one unit.

Let $-\Delta_{m,0} d(\mathbf{S}, \bar{T})$ denote the decrease in distance to the set of feasible base stock levels matrices. For each SKU $m \in M$ and warehouse $n \in N$, let $\mathbf{E}_{m,n}$ be a $|M| \times |N|$ matrix with positions (m', n'), $m' \in M$, $n' \in N$, with ones on positions m and n and zero on all other positions. Then, for a given SKU $m \in M$ of capital good type $c \in C$, we obtain

$$\begin{aligned} \Delta_{m,n} d(\mathbf{S}, \bar{T}) &= d(\mathbf{S} + \mathbf{E}_{m,n}, \bar{T}) - d(\mathbf{S}, \bar{T}) \\ &= \left(EBO_c(\mathbf{S} + \mathbf{E}_{m,n}, \bar{T}) - \mathcal{B}_c^{max} \right)^+ - \left(EBO_c(\mathbf{S}, \bar{T}) - \mathcal{B}_c^{max} \right)^+ . \end{aligned}$$

Increasing the base stock level of a given SKU $m \in M$ at the central warehouse has a decreasing effect on the expected backorders at all local warehouses $n \in N_l$, and no effect on the expected backorders of all other SKUs. Moreover, increasing the base stock level of a given SKU $m \in M$ at some local warehouse $n \in N_l$ has a decreasing effect on the expected backorders for that SKU at that local warehouse and no effect on all other expected backorders. These assertions are easily verified along similar lines as the proof of Lemma 2. It then immediately follows that $-\Delta_{m,n} d(\mathbf{S}, \bar{T}) \geq 0$ for all $m \in M$ and $n \in N$. The ratio $\Gamma_{m,n}^{\mathbf{S}} = \frac{-\Delta_{m,n} d(\mathbf{S}, \bar{T})}{c_a^m}$ denotes the decrease in distance to the set of feasible base stock levels matrices per acquisition cost. During each greedy step, we increase the base stock of SKU m at warehouse n with the highest $\Gamma_{m,n}^{\mathbf{S}}$ to $S_{m,n} + 1$. We continue with these steps until we arrive at a feasible solution $\mathbf{S}$.

6. Computational study

The computational study in this section consists of two parts. In Section 6.1, we report on a case study at NS and present managerial insights. In Section 6.2 and Online Appendix E, we evaluate the performance of both solution approaches and quantify the impact of a dynamic repair policy on total stock investments based on a large test bed of randomly generated instances. We programmed our solution approaches as single threaded applications in C with GLPK as the solver of both linear and integer linear programs. All computations were carried out on a PC running Windows (32 bit) with an Intel Quad Core 2.20 GHz processor and 8 GB RAM.

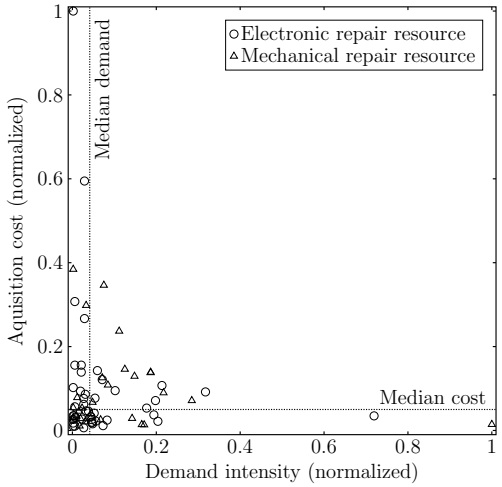
6.1. Case study at NS

NS is the principal passenger railway operator in the Netherlands. Its fleet consists of 900 rolling stock units, divided over twelve different train series. The spare parts inventory system of NS consists of one central warehouse and twelve local warehouses. There is a large repair center incident to the central warehouse. This repair center consists of multiple repair shops, each responsible for a different repair resource.

6.1.1. Setup and objective This case study is focused on the VIRM train series; our case study therefore involves one capital good type, i.e. $|C| = 1$. The VIRM series consist of 176 rolling stock units, all of which are being operated as intercity trains that connect most cities in the Netherlands. We consider the six most important warehouses where the VIRM train series is maintained and leave a handful of locations with only incidental demand out of scope; hence, $|N_l| = 6$.

We select 74 critical SKUs that occur in the configuration of the VIRM series. Of these SKUs, 30 require a mechanical resource for their repair and 44 require an electronic resource for their repair; hence, $|M_r^R| = 2$. The regular and expedited repair lead time for both repair resources is three weeks and one week, respectively. The transportation time is one week and includes administration time and shipment time from the central warehouses to all local warehouses. The acquisition costs of all SKUs range between 150.52 and 23,399.64 euros, and are 2,282.54 euro on average. The total historical demand for all SKUs varies between 1 and 174 per year. In Figure 4, we plot and classify each SKU based on its normalized demand intensity and normalized acquisition cost. This classification will be important when we discuss the results of our case study.

Figure 4 Scatterplot of SKUs in case study



The current practice at NS can be described as follows. On a strategic level, inventory managers decide upon stocking levels using a single-item single-echelon model of the commercial package Servigistics (formerly Xelus Parts Planning). This model does not take into account that NS has the possibility to expedite the repair of parts in short supply. Expediting decisions are then made by inventory managers and repair shop managers together operationally on a weekly basis. For the electronic and mechanical repair shop, we observe from historical data that 30 percent of the total stream of failed parts is expedited.

The main objective of this case study is twofold. First, we want to determine the reductions in investment costs that can be achieved when our solution approaches are used to achieve the same performance as the current approach of NS achieves. Second, and more importantly, we want to understand how a dynamic repair policy can be leveraged to reduce the total investment costs in spare parts while meeting availability targets.

Our benchmark for the case study is the current solution that NS uses. In this solution, the investment in each spare part is determined by the stocking model of Servigistics. For this investment decision, we determine the best achievable availability performance by optimizing expediting decisions and stock placement within our modeling framework. Further details are provided in Online Appendix F. We then use both the DCG algorithm and the greedy heuristic to find an alternative investment decision with at least the same availability performance.

6.1.2. Results and managerial implications Table 2 shows the normalized investment costs of the three approaches to make investment decisions and the corresponding availability performance and expediting fractions. We observe that the greedy heuristic leads to

an investment costs reduction of 52.43% compared to the current approach. As expected, the DCG algorithm has an even higher cost benefit with 53.55%. Apart from the substantial investment costs reductions that can be reaped, it is interesting to note that the gap between the DCG algorithm and the greedy heuristic is small. Later in the numerical experiments of Section 6.2, we will see that this holds across a large variety of industrial size problem instances.

Table 2 Main results case study at NS

Solution approach	Investment costs (normalized)	EBO_{VIRM}	$EXP_{\text{electronic}}$ (%)	$EXP_{\text{mechanical}}$ (%)
Current approach NS	100	13.70	29.82	29.98
Greedy heuristic	47.57	13.65	29.99	29.97
DCG algorithm	46.45	13.65	29.99	29.94

Recall that we classified all SKUs into four distinct SKU groups based on their acquisition costs and demand intensities. To illustrate how the decisions of our new approaches realize the substantial cost reductions reported in Table 2, we will investigate the performance of each of these SKU groups. To facilitate presentation, we first introduce some additional notation. Let G denote the set of different SKU groups, hence $G = \{\text{high demand, low demand}\} \times \{\text{high cost, low cost}\}$ and thus $|G| = 4$. The SKUs that belong to group $g \in G$ are contained in the set M_g^G. For each of the three investment decisions consisting of acquired stock $\mathbf{S}$ and expedite thresholds T, we now calculate the following performance measures:

$$\overline{EXP}_g(T) = \frac{\sum_{m \in M_g^G} EXP_m(T_m)}{|M_g^G|},$$

$$EXP_g(T) = \sum_{r \in R} \sum_{m \in M_g^G} \delta_m^r EXP_m(T_m),$$

$$STOCK_g(\mathbf{S}) = \frac{\sum_{m \in M_g^G} \sum_{n \in N} S_{m,n}}{\sum_{m \in M_g^G} \lambda_{m,0}},$$

$$EBO_g(\mathbf{S}, T) = \sum_{m \in M_g^G} \sum_{n \in N_l} EBO_{m,n}(\mathbf{S}_m, T_m), \text{ and}$$

$$COST_g(\mathbf{S}) = 100 \cdot \frac{\sum_{m \in M_g^G} \sum_{n \in N} c_a^m S_{m,n}}{C(\mathbf{S})},$$

which all provide meaningful information about SKU group $g \in G$. The mean expedited repair utilization and the total aggregate mean expedited repair utilization of g are given by $\overline{EXP}_g(T)$ and $EXP_g(T)$, respectively. Note that $\overline{EXP}_g(T) \in [0, 1]$ and $EXP_g(T) \in [0, 0.6]$.

$STOCK_g(\mathbf{S})$ provides a normalized measure of how much stock of all SKUs in g is acquired. The total mean number of backorders for g is given by $EBO_g(\mathbf{S}, T)$. Finally, $COST_g(\mathbf{S})$ measures the relative difference between the investment costs in g and the overall total costs under the investment decision.

Table 3 provides the performance measures for each SKU group $g \in G$ under each of the three investment decisions. For now, we only consider the performance measures of our solution approaches, and we turn our attention to the left upper quadrant: SKUs with low demand intensities and high acquisition costs. As the table indicates, the unavailability due to this group of SKUs is kept relatively low by providing full repair priority to failed parts rather than by investing in spare parts. Although failed parts always receive expedited repair, this SKU group utilizes only 9.22% of the total available expediting capacity. Conversely, if we look at the right lower quadrant, SKUs with low acquisition costs and high demand intensities receive almost no expedited repair. Instead, the unavailability due to this group of SKUs is kept relatively low by acquiring large amounts of spare parts.

Table 3 Performance measures per SKU group under each solution approach

		Solution approach						
		DCG	Greedy	Current	DCG	Greedy	Current	Measure
Acquisition cost	High	100	100	9.34	67.08	65.56	34.63	$\overline{EXP}$ (%)
		9.22	9.22	1.25	46.73	41.96	27.25	EXP (%)
		0.00	0.00	92.57	15.62	16.17	23.98	$STOCK$
		0.00	0.00	37.75	52.51	54.37	46.18	$COST$
		1.94	1.94	0.92	8.39	8.54	5.76	EBO
	Low	55.57	61.39	13.43	5.11	10.83	36.60	$\overline{EXP}$ (%)
		2.04	4.18	1.26	1.94	4.60	30.04	EXP (%)
		66.00	66.00	104.5	44.24	44.24	23.78	$STOCK$
		5.08	4.22	4.92	42.41	41.41	11.14	$COST$
		1.09	1.05	0.96	2.22	2.12	6.06	EBO
			Low			High		
		Demand intensity						

For the other two SKU groups, our solution approaches neither solely invest in spare parts nor solely expedite the repair of failed parts. If we look at the SKUs with low demand intensities and low acquisition costs, we indeed observe that this group has a large amount of normalized acquired stock as well as a high average expedited repair utilization. As a result, this group has the smallest mean number of backorders of all groups. The impact on the total investment costs and the total available expediting capacity is however small as both demand intensities and acquisition costs are low.

From the right upper quadrant, we observe that a large part of the available expediting capacity is utilized by the group of SKUs with both high demand intensities and high acquisition costs. Although the investment costs in this group are more than half of the total costs of the investment decision, the normalized acquired stock is relatively small. Finally, with more than 8 expected backorders, the unavailability due to this group of SKUs is significantly larger than all other groups.

Our integrated solution approaches thus lead to well-balanced investment decisions in which we acquire large amounts of spare parts of SKUs with low acquisition costs. In doing so, we maximize the availability of these SKUs at relatively low investment costs. Almost all available expediting capacity is then leveraged to dynamically prioritize the repair of failed parts with high acquisition costs, which allows us to refrain from excessively acquiring spare parts with such high costs. The current approach leads to a less balanced investment decision. As Table 3 indicates, the current approach invests heavily in spare parts with high acquisition costs and mainly prioritizes the repair of failed parts of SKUs with high demand intensities.

6.2. Numerical experiments

In the previous section, we have described how our solution approaches leverage a dynamic repair policy to reduce the total investment costs in spare parts while meeting availability constraints. In this section, we assess the value of having such an advanced dynamic repair policy in the first place. We also investigate whether our solution approaches find solutions that are close to optimal and whether they find such solutions within reasonable time.

To answer these questions, we consider a large test bed of 2592 randomly generated problem instances whose parameter values are based on representative data for the capital goods industry. Our test bed consists of both symmetric instances, in which the demand intensities across all local warehouses are identical but varied for different SKUs, and asymmetric instances, in which the demand intensities are varied across all local warehouses and SKUs. For further details regarding the test bed, see Table 5 in Online Appendix E.

To quantify the value of our dynamic repair policy, we create a state-of-the-art 'benchmark' instance for each 'original' instance of problem (P) that we generate. This benchmark instance is identical to the original instance except that it is not possible to differentiate repair lead times through expediting. The mean repair lead time of this benchmark instance is then kept less than or equal to the mean repair lead time of the original instance. This is

achieved as follows: We set $\mathcal{E}_r^{max}$ to 1.0 for each repair resource $r \in R$ in the original instance such that it is feasible (and optimal) to expedite all repairs. We then change the expedited lead time $t_{m,0}^2$ of each SKU $m \in M$ to the shortest mean repair lead time possible in the feasible solution to the original instance. For a given SKU $m \in M$ that requires resource $r \in R$ for its repair, this shortest mean repair lead time is $(1 - \mathcal{E}_r^{max}) \cdot (t_{m,0}^1 + t_{m,0}^2) + \mathcal{E}_r^{max} \cdot t_{m,0}^2$. For this benchmark instance, we compute a lower bound on the optimal cost using the method described in Section 5.1. We denote this lower bound by C_{BENCH}^{LB} and we compare it with C_P^{UB} of the original instance, obtained by the DCG algorithm. That is, $\%RED = 100 \cdot \frac{C_{BENCH}^{LB} - C_P^{UB}}{C_{BENCH}^{LB}}$, where $\%RED$ will indicate how much stock investment reductions can be achieved because of the possibility to expedite the repair of parts in short supply.

To evaluate the effectiveness of our solution approaches, we compute a feasible solution for each generated instance using both solution approaches and we measure the relative difference between the total cost obtained by the solution approach and the corresponding lower bound. That is, $\%GAP = 100 \cdot \frac{C_P^{UB} - C_P^{LB}}{C_P^{LB}}$, where C_P^{LB} is obtained using the method described in Section 5.1 and where C_P^{UB} is obtained using the method described in Section 5.3 in case of the DCG algorithm, or using the method described in Section 5.4 in case of the greedy heuristic.

6.2.1. Results from numerical experiments The main results of our numerical experiments are summarized in Table 4. Detailed results are provided in Online Appendix E. We note that the solutions to the problem instances generally exhibit the same behaviour as extensively described in the case study.

The numerical experiments indicate that both solution approaches perform very well. The average optimality gaps of the DCG algorithm over the asymmetric and the symmetric problem instances are only 0.26 and 0.28, respectively. The optimality gaps of the feasible solutions found by the greedy heuristic are slightly larger. The average optimality gaps of this solution approach are 1.07 and 3.69 over the asymmetric and the symmetric problem instances, respectively. The greedy heuristic is the most efficient heuristic in terms of computation time. Although the computation time of the DCG algorithm is considerably higher, it is still acceptable given the size and strategic nature of the decision problem.

The stock investment reductions that can be achieved because of the possibility to expedite the repair of parts in short supply are quite high with an average stock investment

reduction of around 7.9 percent and even reductions of up to 19.61 percent. Before concluding this section, we briefly return to the case study at NS. The value of a dynamic repair policy in their setting is substantial with a stock investment reduction of 36.40 percent. This is not surprising because our numerical experiments indicate that the value of a dynamic repair policy increases in the additional regular repair lead time or in the fraction of total demand that may be expedited. Both input parameters are slightly larger in the case study than in the problem instances of our test bed.

Table 4 Main results numerical experiments

	DCG algorithm				Greedy heuristic				Benchmark	
	$\%GAP$		CPU time (s)		$\%GAP$		CPU time (s)		$\%RED$	
Problem instances	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
Asymmetric	0.26	0.75	90.08	939.03	1.07	3.15	1.55	11.47	7.95	18.81
Symmetric	0.28	0.77	111.34	1271.85	3.69	8.43	1.75	12.85	7.92	19.61
Total	0.27	0.77	100.71	1271.85	2.38	8.43	1.66	12.85	7.94	19.61

7. Concluding remarks

We have considered a multi-item two-echelon spare parts inventory system, where each warehouse keeps multiple repairable types to maintain several types of capital goods, and where the repair shop at the central warehouse has two options for the repair of each defective part: a regular repair option and an expedited repair option. Irrespective of the repair option, each defective part uses a certain resource for its repair. Assuming a dual-index policy at the central warehouse and base stock control at the local warehouses, we have proposed an exact evaluation procedure for a given control policy.

To find an optimal control policy, we have formulated an optimization problem aimed at minimizing the total investment costs under constraints on both the aggregate mean number of backorders per capital good type and the aggregate mean fraction of repairs that are expedited per repair resource. We have shown how this non-linear non-convex integer programming problem can be decomposed into independent Newsvendor type sub-problems per repairable type, which subsequently allows us to use column generation algorithms. As an alternative solution approach, we have presented an efficient greedy heuristic. Both solution approaches perform very well across a large test bed of industrial size.

We have shown that a dynamic repair policy is effective in reducing the stock investment needed to meet availability requirements for multiple types of capital goods. Based on

a case study at NS, we have shown that our solution approaches lead to well-balanced investment decisions in which large amounts of spare parts of SKUs with low acquisition costs are acquired. In doing so, the availability of these SKUs can be maximized at relatively low investment costs. Almost all available expediting capacity can then be leveraged to dynamically prioritize the repair of failed parts with high acquisition costs, which allows us to refrain from excessively acquiring spare parts with such high costs.

Although our main focus is spare parts inventory systems for repairables, our results are applicable to a wide range of two-echelon distribution systems with a similar structure as in this paper. Particularly the option to expedite the repair of a failed part can be interpreted as a faster, but more expensive second supply source in distribution systems for consumables. In this case, it is natural to charge an additional cost for expediting. We have shown that our results are readily extended to allow for such expediting costs as well.

References

Adan, I.J.B.F., A. Sleptchenko, G.J.J.A.N. Van Houtum. 2009. Reducing costs of spare parts supply systems via static priorities. *Asia-Pacific Journal of Operational Research* **26**(04) 559–585.

Aggarwal, P.K., K. Moinzadeh. 1994. Order expedition in multi-echelon production/distribution systems. *IIE transactions* **26**(2) 86–96.

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

Alvarez, E.M., M.C. Van der Heijden. 2014. On two-echelon inventory systems with poisson demand and lost sales. *European journal of operational research* **235**(1) 334–338.

Alvarez, E.M., M.C. Van der Heijden, W.H.M. Zijm. 2013. The selective use of emergency shipments for service-contract differentiation. *International Journal of Production Economics* **143**(2) 518–526.

Alvarez, E.M., M.C. Van der Heijden, W.H.M. Zijm. 2015. Service differentiation in spare parts supply through dedicated stocks. *Annals of operations research* **231**(1) 283–303.

Arts, J.J. 2017. A multi-item approach to repairable stocking and expediting in a fluctuating demand environment. *European Journal of Operational Research* **256**(1) 102–115.

Arts, J.J., R.J.I. Basten, G.J.J.A.N. Van Houtum. 2016. Repairable stocking and expediting in a fluctuating demand environment: Optimal policy and heuristics. *Operations Research* **64**(6) 1285–1301.

Arts, J.J., M. Van Vuuren, G.P. Kiesmüller. 2011. Efficient optimization of the dual-index policy using markov chains. *IIE Transactions* **43**(8) 604–620.

Basten, R.J.I., G.J.J.A.N. Van Houtum. 2014. System-oriented inventory models for spare parts. *Surveys in operations research and management science* **19**(1) 34–55.

Bitton, S., I. Cohen, M.A. Cohen. 2018. Joint repair sourcing and stocking policies for repairables using erlang-a and erlang-b queueing models Working paper.

Caggiano, K.E., J.A. Muckstadt, J.A. Rappold. 2006. Integrated real-time capacity and inventory allocation for reparable service parts in a two-echelon supply system. *Manufacturing & Service Operations Management* **8**(3) 292–319.

Caglar, D., C.L. Li, D. Simchi-Levi. 2004. Two-echelon spare parts inventory system subject to a service constraint. *IIE Transactions* **36**(7) 655–666.

Cohen, M., P.V. Kamesam, P. Kleindorfer, H. Lee, A. Tekerian. 1990. Optimizer: Ibm's multi-echelon inventory system for managing service logistics. *Interfaces* **20**(1) 65–82.

Cohen, M.A., Y.S. Zheng, V. Agrawal. 1997. Service parts logistics: a benchmark analysis. *IIE transactions* **29**(8) 627–639.

Dada, M. 1992. A two-echelon inventory system with priority shipments. *Management Science* **38**(8) 1140–1153.

Dantzig, G.B., P. Wolfe. 1960. Decomposition principle for linear programs. *Operations research* **8**(1) 101–111.

Drent, M. 2017. Stocking and expediting in two-echelon spare parts inventory systems under system availability constraints. Master's thesis, Eindhoven University of Technology. URL `https://pure.tue.nl/ws/files/75591787/Master_Thesis_Melvin_Drent.pdf`.

Feng, Q., S.P. Sethi, H. Yan, H. Zhang. 2006. Are base-stock policies optimal in inventory problems with multiple delivery modes? *Operations Research* **54**(4) 801–807.

Fischetti, M., F. Glover, A. Lodi. 2005. The feasibility pump. *Mathematical Programming* **104**(1) 91–104.

Gallego, G., Ö. Özer, P. Zipkin. 2007. Bounds, heuristics, and approximations for distribution systems. *Operations Research* **55**(3) 503–517.

Graves, S.C. 1985. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science* **31**(10) 1247–1256.

Gross, J.F., J.M. Thompson, C.M. Harris. 2008. *Fundamentals of Queueing Theory*. 4th ed. Wiley-Interscience, New York, NY, USA.

Howard, C., J. Marklund, T. Tan, I. Rijnen. 2015. Inventory control in a spare parts distribution system with emergency stocks and pipeline information. *Manufacturing & Service Operations Management* **17**(12) 142–156.

Kellerer, H., U. Pferschy, D. Pisinger. 2004. *Knapsack problems.*. Springer, Berlin.

Kranenburg, A.A., G.J.J.A.N. Van Houtum. 2007. Effect of commonality on spare parts provisioning costs for capital goods. *International Journal of Production Economics* **108**(1) 221–227.

Kranenburg, A.A., G.J.J.A.N. Van Houtum. 2009. A new partial pooling structure for spare parts networks. *European Journal of Operational Research* **199**(3) 908–921.

Lam, S. 1977. Queuing networks with population size constraints. *IBM J. Res. Devel.* **21** 370–378.

Lee, H.L. 1987. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science* **33**(10) 1302–1316.

Lübbecke, M.E., J. Desrosiers. 2005. Selected topics in column generation. *Operations Research* **53**(6) 1007–1023.

Messerli, E.J. 1972. Proof of a convexity property of the erlang b formula. *The Bell System Technical Journal* **51**(4) 951–953.

Minner, S. 2003. Multiple-supplier inventory models in supply chain management: A review. *International Journal of Production Economics* **81** 265–279.

Minner, S., E.B. Diks, A.G. De Kok. 2003. A two-echelon inventory system with supply lead time flexibility. *IIE Transactions* **35**(2) 117–129.

Moinzadeh, K., P.K. Aggarwal. 1997. An information based multiechelon inventory system with emergency orders. *Operations Research* **45**(5) 694–701.

Moinzadeh, K., C.P. Schmidt. 1991. An (s-1, s) inventory system with emergency orders. *Operations Research* **39**(2) 308–321.

Muckstadt, J.A. 1973. A model for a multi-item, multi-echelon, multi-indenture inventory system. *Management Science* **20**(4) 472–481.

Muckstadt, J.A. 2005. *Analysis and algorithms for service parts supply chains*. Springer Science & Business Media.

Pyke, D.F. 1990. Priority repair and dispatch policies for reparable-item logistics systems. *Naval Research Logistics (NRL)* **37**(1) 1–30.

Rong, Y., Z. Atan, L.V. Snyder. 2017. Heuristics for basestock levels in multiechelon distribution networks. *Production and Operations Management* **26**(9) 1760–1777.

Rustenburg, W.D. 2000. A system approach to budget-constrained spare parts management. Ph.D. thesis, Eindhoven University of Technology.

Rustenburg, W.D., G.J.J.A.N. Van Houtum, W.H.M. Zijm. 2001. Systeemgericht spare parts management bij de nederlandse koninklijke marine. *Bedrijfskunde* **73**(2) 28–39.

Sheopuri, A., G. Janakiraman, S. Seshadri. 2010. New policies for the stochastic inventory control problem with two supply sources. *Operations Research* **58**(3) 734–745.

Sherbrooke, C.C. 1968. Metric: A multi-echelon technique for recoverable item control. *Operations Research* **16**(1) 122–141.

Sherbrooke, C.C. 2004. *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques*. Kluwer Academic Publishers, Norwell, MA, USA.

Simon, R.M. 1971. Stationary properties of a two-echelon inventory model for low demand items. *Operations Research* **19**(3) 761–773.

Sleptchenko, A., M.C. Van der Heijden, A. Van Harten. 2005. Using repair priorities to reduce stock investment in spare part networks. *European Journal of Operational Research* **163**(3) 733–750.

Song, J.S., P. Zipkin. 2009. Inventories with multiple supply sources and networks of queues with overflow bypasses. *Management Science* **55**(3) 362–372.

Tiemessen, H.G.H., G.J.J.A.N. Van Houtum. 2013. Reducing costs of repairable inventory supply systems via dynamic scheduling. *International Journal of Production Economics* **143**(2) 478–488.

Topan, E., Z.P. Bayındır, T. Tan. 2017. Heuristics for multi-item two-echelon spare parts inventory control subject to aggregate and individual service measures. *European Journal of Operational Research* **256**(1) 126–138.

Van Houtum, G.J.J.A.N., A.A. Kranenburg. 2015. *Spare parts inventory control under system availability constraints*, vol. 227. Springer.

Veeraraghavan, S., A. Scheller-Wolf. 2008. Now or later: A simple policy for effective dual sourcing in capacitated systems. *Operations Research* **56**(4) 850–864.

Whittemore, A.S., S.C. Saunders. 1977. Optimal inventory under stochastic demand with two supply options. *SIAM Journal on Applied Mathematics* **32**(2) 293–305.

Wong, H., A.A. Kranenburg, G.J.J.A.N. Van Houtum, D. Cattrysse. 2007. Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse. *OR spectrum* **29**(4) 699.

Appendix A: Proofs

Proof of Theorem 1 Let $S_{m,0}$ and T_m be fixed. Let $f : \mathbb{N}^{|N_l|} \to \mathbb{R}$ be the part of objective function (12) that depends on $S_{m,n}$, $n \in N_l$. Then, by omitting constants, objective function (12) reduces to

$$f(S_{m,1}, S_{m,2}, \ldots, S_{m,|N_l|}) = \sum_{n \in N_l} \left[c_a^m S_{m,n} - p_c EBO_{m,n}(S_{m,n}) \right],$$

where $EBO_{m,n}$ now depends only on $S_{m,n}$ because $S_{m,0}$ and T_m are fixed. By observing that each term in f is precisely the cost of an independent Newsvendor type problem, one for each local warehouse $n \in N_l$, the desired result directly follows. □

Proof of Lemma 1 Let T_m be fixed and $S_{m,n} = 0$ for all $n \in N_l$. Let $f : \mathbb{N} \to \mathbb{R}$ be the part of objective function (12) that depends on $S_{m,0}$. Then, by omitting constants, objective function (12) reduces to

$$f(S_{m,0}) = c_a^m S_{m,0} - p_c \sum_{n \in N_l} EBO_{m,n}(S_{m,0}),$$

where $EBO_{m,n}(S_{m,0})$ now depends only on $S_{m,0}$ because T_m is fixed and $S_{m,n} = 0$ for all $n \in N_l$.

Recall that the number of parts outstanding at local warehouse $n \in N_l$ is the sum of the demand during transport and shipping time $t_{m,n}$ from the central warehouse to local warehouse n, $D_{m,n}$, and the number of backorders at the central warehouse that belong to local warehouse n. Hence, since $S_{m,n} = 0$ for all $n \in N_l$, $EBO_{m,n}(S_{m,0})$ is equal to the sum of the expected backorders at the central warehouse that are from local warehouse n and the expectation of $D_{m,n}$ (see Equation (2)).

Then, since the number of backorders at the central warehouse that belong to local warehouse n is binomially distributed for a fixed total number of backorders (see Equation (1)), we have

$$f(S_{m,0}) = c_a^m S_{m,0} - p_c \sum_{n \in N_l} \frac{\lambda_{m,n}}{\lambda_{m,0}} EBO_{m,0}(S_{m,0}) = c_a^m S_{m,0} - p_c EBO_{m,0}(S_{m,0}),$$

where we have used the definition of $\lambda_{m,0}$ and the fact that $\mathbb{E}[D_{m,n}]$ is constant and can thus be omitted. By observing that f is precisely the cost of a Newsvendor type problem, the desired result directly follows. □

Proof of Lemma 2 Let T_m and $S_{m,0}$ be fixed and take some local warehouse $n \in N_l$. Let $Y \geq_{st} \tilde{Y}$ denote that a random variable Y is stochastically larger than another random variable $\tilde{Y}$ in the usual stochastic order. Then, observe that $BO_{m,0}(S_{m,0}, T_m) \geq_{st} BO_{m,0}(S_{m,0}+1, T_m)$. This implies that $BO^n_{m,0}(S_{m,0}, T_m) \geq_{st} BO^n_{m,0}(S_{m,0}+1, T_m)$, and thus $BO^n_{m,0}(S_{m,0}) + D_{m,n} \geq_{st} BO^n_{m,0}(S_{m,0}+1) + D_{m,n}$, which is equivalent to $X_{m,n}(S_{m,0}, T_m) \geq_{st} X_{m,n}(S_{m,0}+1, T_m)$. Hence, in particular it holds that $P\{X_{m,n}(S_{m,0}, T_m) \leq x\} \leq P\{X_{m,n}(S_{m,0}+1, T_m) \leq x\}$ for any $x \in \mathbb{N}$. Hence, as $S^*_{m,n}(S_{m,0}, T_m)$ is the smallest $S_{m,n}(S_{m,0}, T_m)$ that satisfies Equation (14), we must have that $S^*_{m,n}(S_{m,0}+1, T_m) \leq S^*_{m,n}(S_{m,0}, T_m)$. □

Proof of Theorem 2 This follows directly from Lemma 1 and Lemma 2. □

Appendix B: Pseudo-code of solution approaches

This section provides pseudo-code of the DCG algorithm, including the exact solution method for $(SUB(m))$, as well as the two-step greedy heuristic. Note that in the pseudo-code of the exact solution method for $(SUB(m))$ (i.e. Algorithm 2), we continue the search over T_m by examining 4 values beyond the last observed local minimum, that is $N^{max} = 4$.

Algorithm 1 DCG algorithm for problem (P)

Step 1: Initialization
  Determine an initial set of trivial policies $K_m^{res} \subseteq K_m$ for each SKU $m \in M$;
Step 2: Master Problem
  Solve the restricted master problem (RMP) $(7)-(11)$ with K_m replaced by K_m^{res};
  Obtain primal and dual solution;
Step 3: Column generation sub-problem
  For the dual variables obtained in Step 2, execute Algorithm 2 for each SKU $m \in M$;
Step 4: Termination test
  If Step 3 results in any policies with negative costs, add these to K_m^{res} and go to Step 2;
  Else solve final version of (RMP) as an integer linear program and obtain a solution for problem (P);

Algorithm 2 Solution method for $(SUB(m))$

Step 1: Initialization
  Set T_m and N to 0, and N^{max} to 4;
Step 2: Initialization per T_m
  Determine upper bound $\bar{S}_{m,0}(T_m)$ using Lemma 1 and set $S_{m,0}$ to 0;
Step 3: While $S_{m,0} \leq \bar{S}_{m,0}(T_m)$
  Determine $S^*_{m,n}(S_{m,0}, T_m)$ using Theorem 1;
  Determine corresponding reduced costs using Equation (12);
  If lowest reduced costs per T_m so far then store policy $(S^*_{m,0}, T_m)$ and corresponding reduced costs;
  Increase $S_{m,n}$ by 1;
Step 4: Termination test
  If reduced costs of $(S^*_{m,0}, T_m) > (S^*_{m,0}, T_m - 1)$ then $N = N + 1$ else $N = 0$;
  If $N \geq N^{max}$ then stop else increase T_m by 1 and go to Step 2;

Algorithm 3 Greedy heuristic for problem (P)

Step 1: Determine vector of expedite thresholds $\bar{T}$
  For each repair resource $r \in R$
    Set T_m to 0 $\forall\ m \in M_r^R$;
    Calculate $\Gamma_m^T\ \forall\ m \in M_r^R$;
    While $d(T) > 0$:
      Determine m' with $\Gamma_{m'}^T \geq \Gamma_m^T\ \forall\ m \in M_r^R$;
      Increase $T_{m'}$ with 1;
      Calculate $\Delta_m d(T)$ and update $\Gamma_m^T\ \forall\ m \in M_r^R$;
    Set $\bar{T}$ to T;
Step 2: Determine matrix of base stock levels $\mathbf{S}$
  For each capital good type $c \in C$
    Set $S_{m,n}$ to 0 $\forall\ m \in M_c^C$, $n \in N$;
    Calculate $\Gamma_{m,n}^{\mathbf{S}}\ \forall\ m \in M_c^C$, $n \in N$;
    While $d(\mathbf{S}, \bar{T}) > 0$:
      Determine (m', n') with $\Gamma_{m',n'}^{\mathbf{S}} \geq \Gamma_{m,n}^{\mathbf{S}}\ \forall\ m \in M_c^C$, $n \in N$;
      Increase $S_{m',n'}$ with 1;
      Calculate $\Delta_{m,n} d(\mathbf{S})$ and update $\Gamma_{m,n}^{\mathbf{S}}\ \forall\ m \in M_c^C$, $n \in N$;

Appendix C: Expediting repairs at additional costs

Rather than imposing a hard constraint on the fraction of repairs that can be expedited per repair resource, we now charge additional costs for expedited repairs. We can model this in two ways. First, we minimize a total initial cost consisting of both the total investment costs in spare parts (as in the original model) and the total expected discounted expediting costs over an infinite horizon. Second, we minimize a total cost rate per time unit consisting of the total depreciation cost rate in spare parts and the total expediting cost rate, where the depreciation cost rate is obtained by depreciating the total initial investment costs in spare parts over the useful life span of these spare parts. Although both cases are distinct from a modeling perspective, they are in fact equivalent from a mathematical point of view and we therefore opt for the second case; see also remark 1 at the end of this section.

Hence, we seek to minimize the total cost rate consisting of the total depreciation cost rate in spare parts and the total repair expediting cost rate while keeping the mean number of aggregate backorders for each capital good type $c \in C$ below $\mathcal{B}_c^{max}$. Let the cost of expediting the repair of one part of SKU $m \in M$ be denoted by c_e^m. The depreciation cost rate of each part of SKU m per time unit, which can be obtained by linearly depreciating the acquisition cost c_a^m of each part of SKU m over its useful life span, is denoted by c_d^m. For a given control policy $(\mathbf{S}, T)$, the total depreciation cost rate in spare parts is defined as $C_d(\mathbf{S}) = \sum_{m\in M}\sum_{n\in N} c_d^m S_{m,n}$ and the total repair expediting cost rate as $C_e(T) = \sum_{m\in M} c_e^m \lambda_{m,0} EXP_m(T_m)$.

The mathematical formulation corresponding to the decision problem described above, which we call problem $(A1)$, is then given as follows:

$$(A1) \qquad \min_{\{\mathbf{S},T\}} \quad C_d(\mathbf{S}) + C_e(T) \tag{16}$$

$$\text{subject to} \quad EBO_c(\mathbf{S}, T) \le \mathcal{B}_c^{max}, \qquad \forall c \in C \tag{17}$$

$$\mathbf{S} \in \mathscr{S}, \quad T \in \mathbb{N}_0^{|M|}. \tag{18}$$

The DCG algorithm can be applied to problem (C) almost immediately. The master problem is now given by problem $(MPA1)$:

$$\begin{aligned}
(MPA1) \quad \min_{\{x_m^k : m\in M, k\in K_m\}} & \sum_{m\in M}\sum_{n\in N}\sum_{k\in K_m} c_d^m S_{m,n}^k x_m^k + \sum_{m\in M}\sum_{k\in K_m} c_e^m \lambda_{m,0} EXP_m(T_m^k) x_m^k \\
\text{subject to} \quad & \sum_{m\in M_c^C}\sum_{n\in N_l}\sum_{k\in K_m} EBO_{m,n}(\mathbf{S}_m^k, T_m^k) x_m^k \le \mathcal{B}_c^{max}, && \forall c \in C \\
& \sum_{k\in K_m} x_m^k = 1, && \forall m \in M \\
& x_m^k \ge 0, && \forall m \in M, \forall k \in K_m
\end{aligned}$$

The corresponding column generation sub-problem for SKU $m \in M_r^R \cap M_c^C$ is then formulated as follows:

$$\begin{aligned}
(SUBA1(m)) \quad \min_{\{(\mathbf{S}_m, T_m)\}} & \sum_{n\in N} c_d^m S_{m,n} + c_e^m \lambda_{m,0} EXP_m(T_m) - p_c \sum_{n\in N_l} EBO_{m,n}(\mathbf{S}_m, T_m) - \upsilon_m \\
\text{subject to} \quad & \mathbf{S}_m \in \mathbb{N}_0^{|N|}, \qquad T_m \in n \in \mathbb{N}_0.
\end{aligned}$$

Note that for a fixed expedite threshold T_m, $SUBA1(m)$ has exactly the same structure as $SUB(m)$. Hence, all properties as well as the exact solution method presented in Section 5.2 also hold for $(SUBA1(m))$.

It is more involved to adapt the greedy heuristic in such a way that it can be applied to problem $(A1)$ because we cannot iteratively increase expedite thresholds until a feasible solution is obtained. It can be shown that for fixed base stock levels at all warehouses, the expected number of backorders of some capital good type $c \in C$ decreases when the expedite threshold of some SKU $m \in M_c^C$ decreases. Hence, an appropriate greedy heuristic for problem $(A1)$ would be as follows. We first set all base stock levels $S_{m,n}$, $m \in M$, $n \in N$, to zero and all expedite thresholds T_m, $m \in M$, sufficiently high so that no repairs are expedited. Subsequently, we iteratively increase $S_{m,n}$ or decrease T_m, whichever leads to the largest decrease in distance to the set of feasible control policy parameters per increase in the total cost rate. We continue with this iterative procedure until we find a control policy $(\mathbf{S}, T)$ that satisfies constraint (17) of problem $(A1)$.

REMARK 1. Rather than minimizing a total cost rate, one might also be interested in minimizing a total initial cost consisting of both the total investment costs in spare parts (as in the original model) and the total expected discounted expediting costs. To this end, let $\beta > 0$ denote the discounting factor. For a given vector of expedite thresholds T, the total expected discounted expediting costs over an infinite horizon is then given by:

$$\tilde{C}_e(T) = \sum_{m \in M} \int_0^\infty e^{-\beta t} c_e^m \lambda_{m,0} EXP_m(T_m)\, \mathrm{d}t = \sum_{m \in M} \frac{1}{\beta} c_e^m \lambda_{m,0} EXP_m(T_m),$$

where the second equality follows from assuming that the system starts in steady state. The subsequent analysis is now identical to the case with a total cost rate, with $C_e(T)$ changed to $\tilde{C}_e(T)$. ◊

Appendix D: Allowing for commonality

In this section, we relax the assumption that M_c^C and M_r^R partition M, that is, the assumption that each SKU $m \in M$ occurs in the configuration of only one capital good type $c \in C$ and uses only one resource $r \in R$ for its repair. We first introduce additional notation to differentiate between demands for the same SKU that stem from different capital good types. We then briefly describe how problem (P) and its solution approaches change when commonality between SKUs is allowed.

Let $\lambda_{m,n,c}$ denote the demand intensity for SKU $m \in M$ at warehouse $n \in N$ originating from capital good type $c \in C$. If SKU m does not occur in the configuration of capital good type c, then $\lambda_{m,n,c} = 0$ by definition. Let the fraction of demands for SKU m at warehouse n that originate from capital good type c over all demands for that SKU at that warehouse be denoted by $\delta_{m,n}^c = \frac{\lambda_{m,n,c}}{\sum_{k \in C} \lambda_{m,n,k}}$.

The aggregate mean number of backorders for each capital good type $c \in C$ is now given by a weighted sum of the mean number of backorders for all SKUs occurring in the configuration of capital good type c, with the fractions $\delta_{m,n}^c$ as weights. That is,

$$EBO_c(\mathbf{S}, T) = \sum_{m \in M_c^C} \sum_{n \in N_l} \delta_{m,n}^c EBO_{m,n}(\mathbf{S}_m, T_m). \tag{19}$$

The definition of the aggregate mean fraction of failed parts that are expedited per repair resource $r \in R$, i.e. $EXP_r(T)$, remains however the same: SKUs now simply contribute to multiple aggregate mean fractions of expedited repairs whenever they require multiple resources for their repair. Hence, with $EBO_c(\mathbf{S}, T)$ now being defined as in Equation (19), we readily generalize our decision problem to the setting where commonality between SKUs is allowed.

The rest of the analysis goes along similar lines as for the setting without commonality. In particular, constraint (8) in the master problem of the DCG algorithm should be reformulated to

$$\sum_{m \in M_c^C} \sum_{n \in N_l} \sum_{k \in K_m} \delta_{m,n}^c EBO_{m,n}(\mathbf{S}_m^k, T_m^k) x_m^k \leq \mathcal{B}_c^{max}, \quad \forall c \in C,$$

which now incorporates our new definition for the aggregate mean number of backorders.

As SKUs may now belong to multiple capital good types and may now use multiple resources for their repair, the column generation sub-problem for SKU $m \in M$ is now formulated as follows:

$$(SUBA2(m)) \quad \min_{\{(\mathbf{S}_m, T_m)\}} \quad \sum_{n \in N} c_a^m S_{m,n} - \sum_{c \in C} \sum_{n \in N_l} p_c \delta_{m,n}^c EBO_{m,n}(\mathbf{S}_m, T_m) - \sum_{r \in R} \rho_r \delta_m^r EXP_m(T_m) - \upsilon_m$$

$$\text{subject to} \qquad \mathbf{S}_m \in \mathbb{N}_0^{|N|}, \qquad T_m \in n \in \mathbb{N}_0.$$

Note that the structure of $(SUBA2(m))$ is identical to $(SUB(m))$. It is therefore readily verified that all properties as well as the exact solution method presented in Section 5.2 also hold for $(SUBA2(m))$, with the critical fraction $\frac{p_c+c_a^m}{p_c}$ in Theorem 1 and Lemma 1 changed to $\frac{\sum_{c\in C}\delta_{m,n}^c p_c+c_a^m}{\sum_{c\in C}\delta_{m,n}^c p_c}$.

In addition to the DCG algorithm, the two-step greedy approach can also be applied almost immediately to the setting where commonality between SKUs is allowed. The only difference is that the decreases in distances $\Delta_m d(T)$ and $\Delta_{m,n} d(\mathbf{S}, \bar{T})$ should now be calculated over multiple repair resources and multiple capital good types, respectively, with the additional note that in calculating $\Delta_{m,n} d(\mathbf{S}, \bar{T})$, we use Equation (19) for the aggregate mean number of backorders for each capital good type $c \in C$.

Appendix E: Numerical experiments

In this section, we report on our numerical experiments. The main objective of these experiments is to examine how both the $\%GAP$ and $\%RED$ are affected by the input parameters of problem (P). To this end, we consider two test beds of large instances based on data representative for the capital goods industry.

The first test bed consists of 1296 instances obtained through all combinations of the parameter values in Table 5. For each instance, we use an uniform distribution $U[0.005, 0.25]$ to generate the demand intensity for each SKU $m \in M$ at all local warehouses $n \in N_l$. Hence, all instances are symmetric in which demand intensities are identical across all local warehouses but varied for different SKUs. Note that in each instance, we assign all SKUs uniformly at random to a repair resource set M_r^R for $r = 1, \ldots, |R|$.

Table 5 Input parameter values for test bed 1

	Input parameter	No. of choices	Values
1	Number of local warehouses, $\|N_l\|$	3	2, 4, 6
2	Number of capital good types, $\|C\|$	2	2, 4
3	Number of repair resources, $\|R\|$	2	2, 4
4	Number of SKUs per capital good type, $\|M_c^C\|$	3	20, 50, 100
5	Lead time from the central warehouse to local warehouse $n \in N_l$ of SKU $m \in M$, $t_{m,n}$	1	1
6	Expedited repair lead time of SKU $m \in M$, $t_{m,0}^2$	2	1, 2
7	Additional regular repair lead time of SKU $m \in M$, $t_{m,0}^1$	2	3, 5
8	Acquisition cost of SKU $m \in M$, c_a^m	1	$U[100, 1000]$
9	Demand intensity for SKU $m \in M$ at each local warehouse $n \in N_l$, $\lambda_{m,n}$	1	$U[0.005, 0.25]$
10	Maximally allowed mean number of backorders over all SKUs $m \in M_c^C$ for capital good type $c \in C$, $\mathcal{B}_c^{max}$	3	$\nu \sum_{m\in M_c^C} \sum_{n\in N_l} \lambda_{m,n}$ for $\nu = 0.04, 0.06, 0.08$
11	Maximally allowed mean fraction of expedited repairs over all SKUs $m \in M_r^R$ that use repair resource $r \in R$ during their repair, $\mathcal{E}_r^{max}$	3	0.05, 0.10, 0.20

In the second test bed, we consider asymmetric cases in which demand intensities are varied across local warehouses and different SKUs. The same uniform distribution $U[0.005, 0.25]$ is used to generate demand intensities for all SKUs. Next, for each SKU $m \in M$, the demand intensity at each local warehouse $n \in N_l$ is determined by multiplying the generated demand rate of this SKU with a factor generated from a second uniform distribution $U[0.5, 1.5]$. The other parameters are set in the same way as for the first test bed and hence, test bed 2 also results in 1296 instances.

The results of test bed 1 and test bed 2 are summarized in Table 6 and Table 7, respectively. In both tables, we present the average and maximum $\%GAP$ and computation times in seconds of both solution

approaches as well as the average and maximum $\%RED$. We first distinguish between subsets of instances with the same value for a specific input parameter of Table 5 and then present the results for all instances.

Table 6 Summary of computational results for test bed 1 (symmetric demand intensities)

		DCG algorithm				Greedy heuristic				Benchmark	
		$\%GAP$		CPU time (s)		$\%GAP$		CPU time (s)		$\%RED$	
Input parameter	Value	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
Number of local warehouses, $\|N_l\|$	2	0.32	0.77	10.22	71.39	4.47	8.43	0.24	0.77	9.69	19.61
	4	0.27	0.56	63.46	288.74	3.58	5.75	1.28	4.42	7.51	14.66
	6	0.25	0.55	260.33	1271.85	3.03	4.77	3.72	12.85	6.58	13.12
Number of capital good types, $\|C\|$	2	0.29	0.77	70.62	611.31	3.66	8.43	1.17	6.49	7.90	19.61
	4	0.28	0.56	152.05	1271.85	3.73	8.04	2.32	12.85	7.95	18.12
Number of repair resources, $\|R\|$	2	0.26	0.54	110.94	1271.85	3.71	8.43	1.75	12.85	7.97	19.61
	4	0.30	0.77	111.74	1135.98	3.68	8.01	1.75	12.78	7.87	18.12
Number of SKUs per capital good type, $\|M_c^C\|$	20	0.37	0.77	39.63	295.95	3.83	8.43	0.61	2.70	7.78	17.91
	50	0.24	0.51	100.45	731.60	3.64	7.64	1.54	6.43	7.97	19.61
	100	0.23	0.50	193.93	1271.85	3.61	6.52	3.09	12.85	8.02	18.37
Expedited repair lead time, $t_{m,0}^2$	1	0.29	0.77	88.26	926.74	4.00	8.43	1.51	9.99	8.91	19.61
	2	0.27	0.69	134.41	1271.85	3.39	7.27	1.99	12.85	6.94	14.91
Additional regular repair lead time, $t_{m,0}^1$	3	0.29	0.77	72.96	682.63	3.86	8.43	1.45	9.39	6.43	13.70
	5	0.28	0.69	149.71	1271.85	3.53	6.70	2.04	12.85	9.42	19.61
Fraction of total demand that may be backordered, ν	0.04	0.27	0.69	110.98	1104.62	3.82	8.43	1.85	12.85	7.89	18.37
	0.06	0.28	0.63	110.69	992.75	3.72	7.64	1.74	11.72	7.92	19.61
	0.08	0.29	0.77	112.33	1271.85	3.54	8.04	1.65	11.69	7.97	18.12
Fraction of total demand that may be expedited, $\mathcal{E}_r^{max}$	0.05	0.28	0.60	113.61	1271.85	3.35	6.09	1.79	12.85	5.30	9.22
	0.10	0.29	0.69	109.28	1135.98	3.58	8.01	1.76	12.43	7.73	14.81
	0.20	0.28	0.77	111.12	1125.12	4.16	8.43	1.68	11.88	10.74	19.61
Total		0.28	0.77	111.34	1271.85	3.69	8.43	1.75	12.85	7.92	19.61

The main observations drawn from both tables can be summarized as follows:

• The DCG algorithm performs very well. In both test beds, the average and maximum $\%GAP$ are at most 0.28 and 0.77 percent, respectively.

• The greedy heuristic performs also very well in test bed 2 with asymmetric demand intensities. The average and maximum $\%GAP$ in this test bed are only 1.07 and 3.15, respectively. The greedy heuristic performs slightly worse in test 1 with symmetric demand intensities: The average $\%GAP$ is 3.69 but instances with 8 or more do occur. This observation is in line with previous research which examined greedy heuristics in multi-item spare parts problems (e.g., Topan et al. 2017). A possible explanation for this slightly worse performance is due to how the second step of the greedy heuristic works. With symmetric demand intensities, we have the property that if in a given iteration the base stock level of a specific SKU is increased at one local warehouse, then also the base stock levels of the same SKU at all other local warehouses are most likely increased in the succeeding iterations. However, in most practical situations in which each local warehouse serves a distinct market with a different demand structure, one will most likely encounter asymmetric demand intensities and hardly ever symmetric demand intensities.

• The average $\%GAP$ of both solution approaches seem to decrease as the instance size (in terms of the number of local warehouses, capital good types and SKUs per capital good type) becomes larger. This

Table 7 Summary of computational results for test bed 2 (asymmetric demand intensities)

		DCG algorithm				Greedy heuristic				Benchmark	
		$\%GAP$		CPU time (s)		$\%GAP$		CPU time (s)		$\%RED$	
Input parameter	Value	Avg	Max	Avg	Max	Avg	Max	Avg	Max	Avg	Max
Number of local warehouses, $\|N_l\|$	2	0.31	0.75	11.27	54.07	1.66	3.15	0.23	0.75	9.63	18.81
	4	0.25	0.63	59.30	292.86	0.88	2.02	1.17	3.85	7.56	14.70
	6	0.21	0.54	199.68	939.03	0.66	1.24	3.26	11.47	6.65	13.27
Number of capital good types, $\|C\|$	2	0.28	0.75	54.22	396.00	1.05	3.15	1.03	5.79	7.82	18.81
	4	0.24	0.53	125.95	939.03	1.08	2.74	2.07	11.47	8.07	18.63
Number of repair resources, $\|R\|$	2	0.23	0.57	90.14	939.03	1.09	3.15	1.55	11.47	8.07	18.81
	4	0.28	0.75	90.03	876.09	1.05	3.15	1.55	11.00	7.83	18.32
Number of SKUs per capital good type, $\|M_c^C\|$	20	0.34	0.75	31.10	328.09	1.14	3.15	0.55	2.25	7.91	18.81
	50	0.24	0.51	76.26	442.99	1.03	2.44	1.36	5.44	8.04	18.63
	100	0.19	0.49	162.90	939.03	1.03	2.62	2.74	11.47	7.89	18.49
Expedited repair lead time, $t_{m,0}^2$	1	0.26	0.75	71.48	723.60	1.14	3.15	1.34	8.37	9.05	18.81
	2	0.25	0.75	108.69	939.03	1.00	3.15	1.76	11.47	6.84	14.46
Additional regular repair lead time, $t_{m,0}^1$	3	0.26	0.75	61.45	531.93	1.10	3.15	1.30	7.90	6.33	14.24
	5	0.25	0.75	118.72	939.03	1.03	2.99	1.80	11.47	9.57	18.81
Fraction of total demand that may be backordered, ν	0.04	0.26	0.75	89.98	939.03	1.03	2.94	1.63	11.47	8.07	18.32
	0.06	0.25	0.66	90.28	906.21	1.07	2.99	1.55	10.75	7.81	18.81
	0.08	0.26	0.75	90.00	935.16	1.10	3.15	1.48	10.41	7.97	18.63
Fraction of total demand that may be expedited, $\mathcal{E}_r^{max}$	0.05	0.25	0.62	95.22	939.03	0.91	3.15	1.60	11.00	5.12	9.83
	0.10	0.26	0.66	88.70	935.16	1.00	2.94	1.56	11.47	7.84	13.53
	0.20	0.26	0.75	86.32	876.09	1.29	3.15	1.49	10.47	10.88	18.81
Total		0.26	0.75	90.08	939.03	1.07	3.15	1.55	11.47	7.95	18.81

is very convenient since we typically face large-sized instances in practice. The average $\%GAP$ percent of the DCG algorithm tends to increase with the number of repair resources. This is not surprising, because problem (MP) has $|M|+|C|+|R|$ constraints and the same number of basic variables in an optimal solution. Since constraint (10) assures that for each SKU $m \in M$ a convex combination of policies is chosen, there is a basic variable for each SKU m. Hence, there are at most $|C|+|R|$ SKUs for which the optimal solution to problem (MP) is fractional. This explains why the GAP percent increases with the number of repair resources. Note that this does not hold for the number of capital good types because the number of basic variables that increase with the number of capital good types is clearly more than the corresponding increase in the maximum number of SKUs for which the optimal solution to problem (MP) is fractional.

- The average $\%GAP$ of the DCG algorithm tends to decrease as the fraction of total demand that may be expedited or backordered decreases. This also seems to hold for the greedy heuristic, except with symmetric demand intensities: The average $\%GAP$ of the greedy heuristic in test bed 1 increases when the fraction of total demand that may be backordered decreases.

- The greedy heuristic is the most efficient heuristic in terms of computation time. The computation time of the DCG algorithm is considerably higher. Over 98 percent of that computation time is spent on solving the sub-problems. This task can also be parallelized using a multi-threaded approach, which would reduce the computation time of the DCG algorithm even further. The computation time of both solution approaches increases as the problem size (in terms of the number of local warehouses, capital good types and SKUs per capital good type) gets larger and decreases when the means of the repair lead times get smaller.

- The stock investment reductions that can be achieved because of the possibility to expedite the repair of parts in short supply are quite high with an average stock investment reduction of around 7.9 percent and even reductions of up to 19.61 percent.
- The stock investment reductions due to our dynamic repair policy increase when the additional regular repair lead time or the fraction of total demand that may be expedited increase, and decrease when the expedited repair lead time increase.

Appendix F: Approach to determine benchmark for case study

Our benchmark for the case study is the current solution that NS uses. In this solution, the investment in each spare part is determined by the stocking model of Servigistics. Let S_m^{ns} denote the amount of stock of SKU $m \in M$ determined by this stocking model. Given S_m^{ns}, we determine the best achievable availability performance by optimizing expediting decisions and stock placement within our modeling framework. That is, we want to determine an expedite threshold $\tilde{T}_m \in \mathbb{N}_0$ and a base stock levels vector $\tilde{\mathbf{S}}_m \in \left\{\mathbb{N}_0^{|N|} : \mathbf{1} \cdot \mathbb{N}_0^{|N|} = S_m^{ns}\right\}$ such that $EBO(\tilde{\mathbf{S}}, \tilde{T})$ is minimized while $EXP_{\text{mechanical}}(\tilde{T}) \leq 0.3$ and $EXP_{\text{electronic}}(\tilde{T}) \leq 0.3$. This results in the following mathematical formulation of the optimization problem:

$$
\begin{aligned}
(A2) \qquad \min_{\{\tilde{\mathbf{S}}, \tilde{T}\}} \quad & EBO(\tilde{\mathbf{S}}, \tilde{T}) && (20)\\
\text{subject to} \quad & EXP_{\text{mechanical}}(\tilde{T}) \leq 0.3, && (21)\\
& EXP_{\text{electronic}}(\tilde{T}) \leq 0.3, && (22)\\
& \tilde{\mathbf{S}} \in \mathscr{C}, \quad \tilde{T} \in \mathbb{N}_0^{|M|}, && (23)
\end{aligned}
$$

where $\mathscr{C} = \left\{\tilde{\mathbf{S}} : \mathbf{1} \cdot \tilde{\mathbf{S}}_m = S_m^{ns} \ \forall m \in M\right\}$. We solve problem $(A2)$ using a decomposition and column generation approach very much similar to our approach to solve problem (P), described in Section 5. The resulting corresponding sub-problem, however, does not allow for an easy solution method other than enumeration over $\tilde{T}_m$, and for each $\tilde{T}_m$, enumerating over all possible allocations of $\tilde{\mathbf{S}}_m$ over all local warehouses; see Drent (2017) for further details.