Surveys in Operations Research and Management Science 19 (2014) 34–55

Contents lists available at ScienceDirect

Surveys in Operations Research and Management Science

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

Review

System-oriented inventory models for spare parts

R.J.I. Basten [a,*], G.J. van Houtum [b]

[a] University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands

[b] Eindhoven University of Technology, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands

HIGHLIGHTS

- We survey the literature on models for spare parts inventory control.
- Our focus is on models using system-oriented service measures.
- We link the models to two archetypical types of spare parts networks in practice.
- Both the single-location and multi-echelons models are treated.
- We discuss various extensions, including the use of lateral and emergency shipments.

ARTICLE INFO

Article history:
Received 14 July 2013
Received in revised form
15 May 2014
Accepted 19 May 2014

ABSTRACT

Stocks of spare parts, located at appropriate locations, can prevent long downtimes of technical systems that are used in the primary processes of their users. Since such downtimes are typically very expensive, generally system-oriented service measures are used in spare parts inventory control. Examples of such measures are system availability and the expected number of backorders over all spare parts. This is one of the key characteristics that distinguishes such inventory control from other fields of inventory control. In this paper, we survey models for spare parts inventory control under system-oriented service constraints. We link those models to two archetypical types of spare parts networks: networks of users who maintain their own systems, for instance in the military world, and networks of original equipment manufacturers who service the installed base of products that they have sold. We describe the characteristics of these networks and refer back to them throughout the survey. Our aim is to bring structure into the large body of related literature and to refer to the most important papers. We discuss both the single location and multi-echelon models. We further focus on the use of lateral and emergency shipments, and we refer to other extensions and the coupling of spare parts inventory control models to related problems, such as repair shop capacity planning. We conclude with a short discussion of application of these models in practice.

Contents

* Corresponding author. Tel.: +31 53 489 4007.
E-mail addresses: r.basten@utwente.nl (R.J.I. Basten), g.j.v.houtum@tue.nl (G.J. van Houtum).

http://dx.doi.org/10.1016/j.sorms.2014.05.002
1876-7354/

1. Introduction

In this survey, we discuss models and literature on spare parts inventory control. We focus on spare parts inventories for technical systems that are used in the primary processes of their users. Examples are trains, radar systems, MRI-scanners, wafer steppers, and baggage handling systems. Upon failure of such a system, tests are performed to isolate the failure to a specific Line Replaceable Unit (LRU) and this LRU is then replaced by a functioning spare part. This repair-by-replacement policy enables quick repair of the technical system so that the disruption of the primary process of the user is kept within certain limits. This is important, since such disruptions can be very costly; for instance, in the semiconductor industry, downtime costs of the bottleneck machines are estimated to be tens of thousands of euros per hour [1, p. 17]. Obviously, having adequate numbers of spare parts is of key importance for this repair-by-replacement policy to be effective. However, spare parts stocks may tie up a lot of capital: commercial airlines are estimated to have over $40 billion worth of spare parts [2, p. 78], a single company such as ASML, which builds lithography equipment used in semiconductor manufacturing, owns spare parts worth tens of millions of euros [1, p. 78], and the US Coast Guard Aircraft owns inventories worth over $700 million [3, p. 1028]. Stocking the right number of spare parts, not too few and not too many, is thus of key importance. However, stocking the right amounts is difficult, especially for expensive components that fail infrequently and have a long replenishment lead time. In the military world, for instance, lead times can be over a year [4, p. 17].

Spare parts are either repairable or consumable and they differ greatly in their values. In their benchmark study of after-sales service logistics systems, Cohen et al. [5, p. 630] report that the "…average part cost is $270, with some companies reporting parts that cost hundreds of thousands of dollars". Still, the impact of unavailability of a low value spare part and a high value spare part may be the same. Consider a bearing and an X-ray tube, both of which are used in a fully automated security check point in a baggage handling system. If either one of them breaks down, the check point cannot be used anymore. Since the user of the baggage handling system is interested only in whether or not the system is working, it makes sense to stock relatively more inexpensive bearings than expensive X-ray tubes. Due to the direct link between the availability of spare parts and the availability of the technical systems, it makes sense for many companies to use *system-oriented* service measures and targets. Targets can be set, for example, for the availability of the technical systems or the expected number of backorders over all LRUs (a backorder being a spare part that is requested but not yet delivered). This is a key difference with standard inventory models, in which *item-oriented* service measures, such as the fill rate (the percentage of requested parts that can be delivered from stock immediately), are used. In a comparison of multi-item spare parts inventory models (using a system approach) with single-item models, in a single-location setting, Thonemann et al. [6] show that costs savings of about 10%–20% are possible when using a system approach instead of an item approach. Such savings may be achieved when there are large cost differences between the various components. Rustenburg et al. [7,8] study spare parts models for a two-echelon network at the Royal Netherlands Navy and compare the item approach that was in use at that time with the system approach. Rustenburg et al. [8, p. 172] show that for one system, spare parts holding costs would reduce by about 60% under the system approach, while attaining a slightly higher spare parts availability; for another system, the spare parts holding costs would reduce by 9%, while bringing the availability up from 56% to 90%. In our survey, we focus on the system approach and we will discuss the commonly used greedy algorithm that can be used to solve such multi-item models.

Table 1
Overview of the topics of the core chapters.

	Without emergency or lateral shipments	With emergency (and lateral) shipments
Single-location model	Section 3	Section 5.1
Two-echelon model	Section 4	Section 5.2

Technical systems have a long life span. Thales' radar systems, for instance, have a life span of 15–30 years [9] and Lockheed Martin's F-16 Fighting Falcon has a life span of about 25 years [10, p. 3]. During such long life span of a system, three phases can be distinguished (see, e.g., [11]). Spare parts stocks are built up as the installed base is built up in the first year: the initial phase. Next, the spare parts network is in a long maturity phase, in which demand rates are relatively stable (because the size of the installed base remains at a stable level) and there are no problems with the procurement or repair of parts. After that (say, halfway through the life span), the final or end-of-life phase starts: procurement of parts is no longer possible, since suppliers do not have the equipment or materials anymore to manufacture them. Repairs may also be more difficult, for similar reasons. Just before that time, last (time) buys need to be made, or alternative solutions need to be found during the end-of-life phase (see, e.g., [12,13], and the references therein). In parallel or some time later, the size of the installed base gradually decreases because users buy new systems and the old systems are disposed of. In this survey, we focus on spare parts model that can be used in the maturity phase. Furthermore, we focus on models that can be used for spare parts planning at the tactical decision level.

The models that we discuss can be used for inventory control of spare parts that are used for corrective maintenance of random failures. For the forecasting of demand for such spare parts, we refer to the reviewed literature in [14,15]. We will discuss extensions of models for inventory control that incorporate advance demand information, which are relevant under preventive and condition-based maintenance (in Section 6). We discuss models that can be used both by users maintaining their own equipment and by Original Equipment Manufacturers (OEMs) or system integrators that service the equipment that they have sold. We discuss both cases in more detail in Section 2 and we focus in our review on spare parts models that can actually be used in practice.

Because of the importance of and difficulties related to spare parts inventory control under system-oriented service constraints, there has been a lot of research on this topic, starting with the seminal paper of Sherbrooke [16]. We aim to give an up-to-date overview of this large body of literature and we will refer to the most important papers; it is not possible to be complete and this is not what we aim for. For two older extensive overviews of models for spare parts inventory control, we refer to the books of Sherbrooke [17,18][1] and Muckstadt [19].

The remainder of this paper is organized as follows. In Section 2, we discuss real-life networks and their characteristics. Next, four key models are discussed in the core of this paper, see also Table 1. The most simple model, the single-location model, is presented in Section 3. In Section 4, we discuss the two-echelon model, or METRIC model, including the greedy algorithm that can be used to optimize the stock levels. These two models are well known in the literature on spare parts inventories; we extensively survey existing results with the aim of providing a good starting point to researchers and practitioners new to the field and a reference to those that are more experienced in this field. Next, we discuss both the single-location model with emergency shipments and the two-echelon model with lateral and emergency shipments in Section 5. Subsequently, several extensions are discussed in Section 6. In these sections, we survey the relevant literature, ranging from the first results to the latest contributions. Throughout this paper, we refer back to the practical settings of Section 2, in order to show the practical relevance of the models. Finally, in Section 7, we discuss applications of these models in practice and we conclude.

2. Real-life networks

As mentioned in the introduction (Section 1), there are two (extreme) types of networks for which spare parts models are useful: networks of users who maintain the technical systems that they use, and networks of OEMs or system integrators who service the technical systems that they have sold. We refer to the latter type of networks as *OEM networks*, while we call the former *user networks*.

The user networks could be denoted as the more traditional networks. In the past, it was common for users of technical systems to maintain those systems themselves. Later on, OEMs or other parties gradually took over more and more maintenance activities in many industries. In some cases, all maintenance is carried out by the OEM, or the user does not even buy the system anymore, but just the function (including system availability guarantees): 'power by the hour' (in commercial industry, typically airlines, see, e.g., [20]). A main driver for this shift in the past ten to twenty years is the strong increase in technical complexity of equipment and thus of its maintenance. Since an OEM has developed the technical system, it can generally deal much easier with the system's complexity than individual users. Another main driver is that in many industries, companies have implemented lean operations programs, implying that they have smaller buffers for disturbances and thus that they require higher system availabilities for their critical systems. OEMs are usually better able to realize those higher system availabilities, e.g., because they can collect more statistics on failure behavior and they can share spare parts and other maintenance resources for multiple users. Other drivers are, for instance, standardization of equipment (or modules), a possible focus on core business at users, the presence of OEMs who see the maintenance services as an opportunity to gain market share, and technical possibilities to monitor equipment remotely. Obviously, there are also factors, including cultural and political factors, that may block or slow down this shift. For more details on these trends and the various levels of outsourcing, see [21,22]. The various levels of outsourcing imply that in practice, we also see networks with a mix of the features of the user and OEM networks.

In the military industry, most maintenance is still performed by military organizations themselves (at least in Europe), despite the increased complexity of the equipment. This is partly due to the believe that a military organization should not be dependent on civilians. So, a typical example of a user network is the spare parts network of a military organization; the spare parts networks in the high-tech industry (e.g., computers, printing equipment, medical systems) provide typical examples of OEM networks. An overview of the main characteristics of the two archetypes is given in Table 2. We discuss the two archetypes in detail in Sections 2.1 and 2.2, respectively.

2.1. User networks

Examples of user networks have been reported in multiple works in the literature, see Table 3. There are many examples from the military world; other examples are in aviation and railways. For example, KLM (Royal Dutch Airlines) has its own maintenance

[1] In the remainder, we will refer to [18] only, when usually the reference could also have been to [17].

Table 2
Characteristics of two network archetypes.

User networks (typical for military systems)	OEM networks (typical for high-tech systems)
Preventive maintenance dominates	Corrective maintenance dominates
Two echelon levels in one region	Global network with two echelon levels
No emergency option	Emergency option at highest echelon level
Repairs at own repair shops	Repairs at original equipment manufacturers
Relatively loose service targets	Strict/high service targets

Table 3
Examples of users maintaining their own systems.

User	Source
An electric utility company	[23]
Italian airports	[24]
KLM engineering & maintenance	Our experience (e.g., [25])
NedTrain	Our experience (e.g., [26])
Italian paper-making industry	[27]
Royal Netherlands Navy	[28,29,7]
US Air Force	[16,18,19]
US Coast Guard	[3]

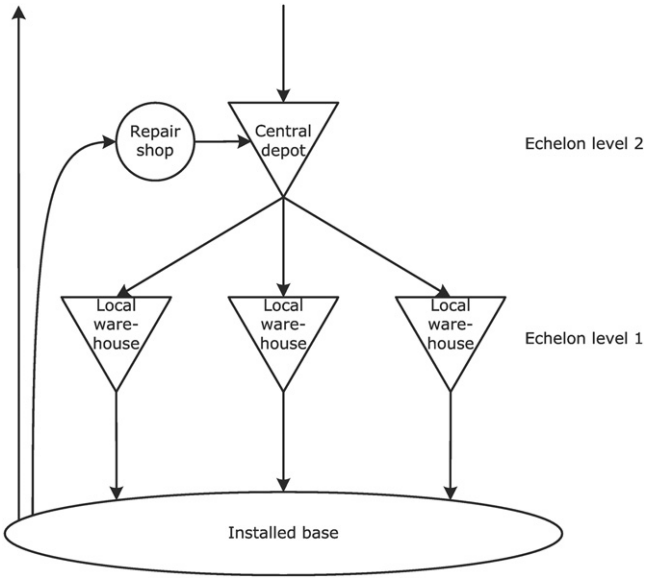

Fig. 1. Archetypical network of a user that maintains its own systems.

organization, KLM Engineering & Maintenance, just as the Netherlands Railways (NS) has NedTrain as its maintenance organization.

Before we discuss the typical network layout and various characteristics of the inventory control in such a network, we focus on the characteristics of the demands for spare parts that are typical in user networks. For an extensive overview of spare parts planning and control for users/maintenance organizations, we refer to [30].

The systems maintained by the companies listed in Table 3 have in common that extensive preventive maintenance programs are used for them. Still, some corrective maintenance is also required. Furthermore, every now and then modifications are applied (e.g., NedTrain typically performs an overhaul halfway through the life of a train, see [31]). This results in both planned and unplanned demands for spare parts. For planned demands, parts can be provided at the moment that they are needed, i.e., it is not necessary to keep safety stock for those demands. Unplanned, random demands can realistically be modeled as a Poisson process (see, e.g., [18, p. 21]).

Planned demands result from modifications and from preventive maintenance. (For an overview of the various preventive maintenance concepts, such as time-based and condition-based maintenance, we refer to [32]). In addition, planned demands can also result from some random failures. Upon such failure, it depends on the criticality of the part and on the design of the technical system whether or not the part should be replaced immediately. For example, if a toilet in a train breaks down, it is usually not necessary to get the train into the repair shop immediately; the toilet is not critical and its replacement can wait until the next regular maintenance visit. Other parts are built in redundantly so that failure of one component does not mean that the technical system stops functioning. Two parts can be built in in parallel, or, more generally, n parts are built in, out of which k ($k < n$) need to function: a k-out-of-n system. Modern radar systems at Thales Nederland, for instance, are equipped with n transmitter–receiver units in one plane, of which only k need to function for the radar to function [33].

Unplanned demands may occur due to a range of reasons. It can be an explicit decision to use the run-to-failure policy and not to perform any preventive maintenance. In other cases, preventive maintenance actions are scheduled too late. This can happen since cost considerations usually lead to scheduling preventive maintenance such that there is a certain probability that parts will fail before the scheduled maintenance action takes place. A next possible reason for occurrence of an unplanned demand, is that for some preventive maintenance and modifications, the planning horizon is shorter than the supply or repair lead time. This is often the case for more expensive parts that are used infrequently; lead times for such items in the military world can be more than a year. Another problem is that some parts are used only in a certain percentage of the preventive maintenance actions: when a module is overhauled, several parts are inspected and replaced only if they do not meet certain criteria, e.g., a train wheel that has become too thin. These parts are called x% parts by some authors [30, p. 12] and this type of maintenance is usually called opportunity-based maintenance or inspection-based maintenance (some authors include it in the term condition-based maintenance). If inspections are performed regularly and x is low, then demands for these spare parts are effectively random. A final reason for seeing unplanned demands for spare parts, is that some companies have one department that is responsible for the maintenance planning, whereas another department is responsible for the spare parts supply (see, e.g., [30, p. 1]). If good communication between these two departments is lacking or if the maintenance planning department regularly changes the planning, demands are random from a spare parts inventory control point of view.

The spare parts network of a user/maintenance organization usually consists of two echelon levels, with the main depot being located at or close to the central repair shop; see Fig. 1 for a typical example. The more expensive components are usually repaired by the maintenance organization or by an outside repair shop, which may be owned by the OEM. Less expensive components are typically not economically repairable and are thus consumable. Less expensive consumables are usually not specific and can be obtained through multiple sources, but other consumables may be sourced at the OEM only. In that case, the lead time can be high, as mentioned above, and possibly unreliable. Repair times at the internal repair shop are usually more reliable.

If outside suppliers supply a few different parts only, then it often makes sense to order those in batches, especially if lead times are high and the components are relatively inexpensive. This will lead to lower ordering costs, since those costs typically have a fixed component per order, irrespective of the order size. Internally in the network, batching per part type makes little sense, since every couple of days usually a shipment of various parts is sent from the central depot to each of the local warehouses. Batching or,

Table 4
Examples of OEMs maintaining sold systems.

OEM	Source
ASML	[1,34,35]
Cisco	Our experience (e.g., [36])
IBM	[37–39]
Océ	Our experience (e.g., [40,41])
Teradyne	[42]
Vanderlande Industries	Our experience (e.g., [43])
Volvo Parts Corporation	[44]

better, consolidation of different parts then occurs automatically. Similarly, it also does not make sense to consider multiple supply modes (transportation modes) in the own network, although that does make sense in some cases at the external supplier (e.g., an expedited repair mode).

For the more expensive parts, there usually exists no emergency option in these networks in the sense that a lead time of less than a day or a few days is achievable. However, in case of an emergency, there typically exists some way to get a part relatively fast. For example, the repair shop can be told to give priority to some repairs, and in aviation, for instance, there usually exist possibilities to get parts from other airlines. Still, usage of these options is to be prevented. Therefore, at the tactical planning level, spare parts are planned as if there is no emergency option whatsoever and demands are backordered.

A further complication in spare parts inventory control for expensive, repairable parts is that repair of a Line Replaceable Unit (LRU) is in many cases achieved by replacing one or more Shop Replaceable Units (SRUs). These, in turn, may also be repaired by replacing smaller components. This means that multi-indenture spare parts models are required.

In many cases, especially in the military world, the service level target is based on the number of backorders. In other cases, (part or order) fill rates are used to measure the service level, although it is acknowledged that there should be a focus on system-oriented service levels (as argued in Section 1). Time-based fill rates are not very common (the percentage of demands that is fulfilled in less than a certain time). Service level targets in user networks are typically not as strict as in OEM networks.

2.2. OEM networks

We give examples of OEMs that provide maintenance services in Table 4. The 14 companies that participated in the aforementioned benchmark study on which Cohen et al. [5] report, are also all OEMs. Interestingly, Cohen et al. report that most companies use a time-based fill rate, i.e., the percentage of components that can be delivered upon request in less than a certain amount of time (say, 2 hours). Since then, the focus has shifted and users are increasingly focusing on the availability of their technical systems. Some users even demand Performance-Based Logistics (PBL, in defense) or 'power by the hour', as mentioned above. OEMs are increasingly interested in providing more after sales service, for two reasons: the first is that customers increasingly require good after sales service so that providing such service gives a competitive advantage, and the second reason is that selling services is at least as profitable as selling goods (for both reasons, see, e.g., [45,46,21,22]). As a result, in many industries, OEMs are, or are increasingly becoming, responsible for after sales service.

Typically, a lot of corrective maintenance is performed, although this is gradually changing in favor of preventive maintenance, in particular condition-based maintenance. The OEM replaces the failed component and, in the case of a more expensive component, sends the component to the repair shop of the company that originally manufactured that component. This could be the OEM itself or another company. In the latter case, the OEM (in the terminology that we have used above) is in fact a system integrator (the term OEM would then better fit the companies that manufacture each of the components).

Fig. 2 gives a typical example of a network as it is used by OEMs that maintain the systems that they have sold. According to Cohen et al. [5, pp. 630–631], "…the three-echelon structure is most prevalent. It is followed in popularity by the two-echelon structure. In most of the three-echelon structures, the middle echelon is dedicated to making emergency shipments only. Therefore, for replenishment purposes, most of the companies use only two echelons". We also usually see two echelon levels for replenishment purposes; emergency shipments are, in our experience, usually supplied from the central depot at one of the three regions or continents. However, there may also be quick response stocks that are located in between the depicted first and second echelon level (e.g., at Océ and Volvo, see [47,48] respectively), or some of the local warehouses act as main warehouses that can perform lateral transshipments to another local warehouse (e.g., at ASML, see [34]). The three central depots at each of the three regions can also usually perform lateral transshipments to each other. In some cases, one of them is a main central depot, for instance, when the manufacturing facility is close to that depot. Having these types of emergency and lateral transshipment options is a key difference with the companies that maintain their own technical systems, partly due to tighter service level targets that are typically used in OEM networks compared with user networks. Because of the various options, it can be relevant to use multiple supply mode models that consider a cheaper, slower supply mode and a faster, more expensive one.

In some cases, there are third parties, instead of the OEM, that take the responsibility upon them to perform maintenance at multiple companies. In some ways, this is not different from OEMs servicing the technical systems. For instance, service contracts are used in both cases. In other respects, however, this is a key difference. For example, if an OEM is responsible for the availability of the technical systems that it manufactures for an extensive period, it makes sense to improve the systems by designing-for-maintenance or designing-for-life-cycle-costs. It also means that the OEM is enabled to do so, because it gets direct feedback from its installed base. This is not the case if a third party performs the maintenance.

3. Single-location model

In this section, we describe single-location, multi-item inventory models. Such models can be used for both user and OEM networks. They can be applied directly in the special case that a network consists of only one stockpoint (which could occur if all technical systems can be supplied fast enough from one stockpoint, e.g., because they are all installed in a relatively small region). This is the situation that is assumed for the description of the basic model in Section 3.1. For larger networks, a single-location model may be a useful building block, e.g., in the following cases:

- In a user or OEM network where the central depot usually has sufficient stock, e.g., because the same parts are still used to manufacture new technical systems (in the case of an OEM network), and where lateral transshipments are seldomly used. It can then be appropriate to determine the base stock levels of each local warehouse via a single-location model, taking into account a certain average delay for the replenishments at the central depot.
- In a user network in which the central depot and the local warehouses are at close distance of each other and where shipments between all stockpoints are possible against a low cost. Such a network can be seen as one large virtual stockpoint for which

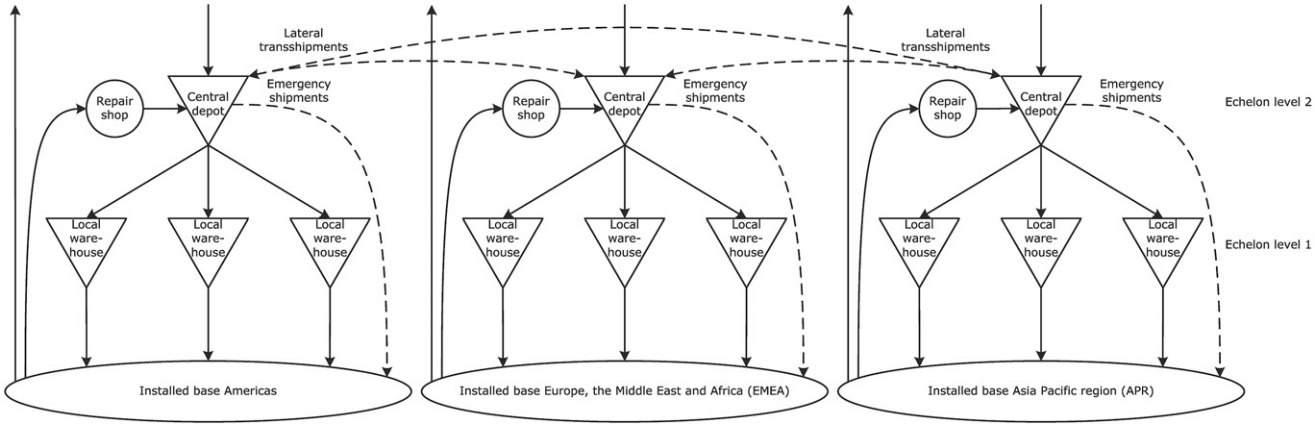

Fig. 2. Archetypical network of an OEM that maintains its sold systems.

the total inventory is determined first, and only in a second stage the allocation of the inventory to the various locations is considered. The single-location model that we discuss below can then be used for the first step.

- In an OEM network in which a certain region has no interaction with the other regions and where all stock points in that region are at close distance of each other. The single-location model can be used for such a region in a similar way as described at the previous bullet.
- In a user or OEM network in which the local warehouses are divided into groups, with lateral transshipments being applied only within each group. Such a network may be modeled as follows. For each group, a multi-location model with lateral transshipments is used, as described in Section 5. For the central warehouse, a single-location model is used, with demand streams resulting from each of the groups.

A single-item version of the basic model of this section was formulated for the first time by Feeney and Sherbrooke [49], who also discussed some extensions (to compound Poisson demand processes and the lost sales case). Sherbrooke [16] extended this model to a multi-item distribution system with one central depot and multiple local warehouse, the METRIC model (see Section 4), and he proposed a greedy heuristic for the determination of base stock levels. Here, we consider a model that generalizes the model of Feeney and Sherbrooke [49] to a multi-item model, and we present a greedy algorithm as one of the possible optimization methods (this model, including its optimization by the greedy algorithm, was first treated by Sherbrooke [17, Chapter 2]). We discuss this model quite extensively because a lot of aspects of more complicated models can already be explained here. This helps the discussions in the remaining sections.

We start with the description of the model with a service level constraint in terms of the aggregate expected number of backordered demands. For the optimization within this basic model we will use a greedy heuristic. The description of the basic model, an explanation of the main assumptions, the evaluation of the model, and its optimization are given in Sections 3.1–3.4, respectively. After that, we discuss alternative optimization techniques in Section 3.5 and alternative service measures in Section 3.6.

3.1. Model description

Consider a single warehouse where several spare parts are kept on stock to serve an installed base of machines of the same type. The machines consist of multiple parts, of which we consider only the critical ones. A failure of a critical part implies that the whole machine goes down. Upon failure, the defective part is taken out of the machine and immediately sent to a repair shop. The machine is repaired by placing a functioning spare part, as soon as one is available. Due to this repair-by-replacement policy, these types of parts are called Line Replaceable Units (LRUs). Defective parts can also be scrapped instead of repaired. In that case, immediately a new part is ordered at an outside supplier. From a modeling point of view, both cases are identical. In our explanation below, however, we will use the terminology of repairable parts.

The set of LRUs is denoted by I, and the number of LRUs is $|I| \in \mathbb{N} := \{1, 2, \ldots\}$. For notational convenience, the LRUs are assumed to be numbered $i = 1, \ldots, |I|$. For each LRU $i \in I$, demands occur according to a Poisson process with a constant rate m_i (≥ 0). The rate m_i denotes the demand rate for all machines together. The total demand rate for all LRUs together is denoted by $M = \sum_{i \in I} m_i$ and we assume that $M > 0$. A demand is fulfilled immediately if possible, and otherwise backordered and fulfilled as soon as possible. Each demand is accompanied by the return of a defective part, which is immediately sent into repair. The time that a defective part spends in the repair shop is called the repair leadtime. Repair leadtimes of different LRUs are assumed to be independent and repair leadtimes of parts of the same LRUs are assumed to be independent and identically distributed (i.i.d.). The mean repair leadtime for LRU i is denoted by t_i (>0). Because each defective part is immediately sent into repair, the inventory position of LRU i, defined as the physical stock minus backordered demand plus parts in repair, is constant. This constant amount is denoted by $S_i \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

Instead of saying that each defective part is immediately sent into the repair shop (or that for each defective part immediately a replacement is ordered), we, more formally, say that for each LRU the stock is controlled by a continuous-review basestock policy with basestock level S_i for LRU i, also written as an $(S-1, S)$ policy or called a one-for-one replenishment policy. The basestock level, which also denotes the initial stocking level, is a decision variable.

The objective is to minimize the total inventory holding costs, subject to a constraint on the aggregate expected number of backorders. Minimizing the investment in spare parts is equivalent to minimizing the inventory holding costs, assuming that both cost types are linear in the number of parts. The inventory holding costs per part of LRU i per time unit is c_i^{h} (>0), the total inventory holding costs for LRU i are then $C_i(S_i) = c_i^{\mathrm{h}} S_i$, and the aggregate inventory holding costs are given by:

$$C(\mathbf{S}) = \sum_{i \in I} C_i(S_i) = \sum_{i \in I} c_i^{\mathrm{h}} S_i,$$

where $\mathbf{S} = (S_1, \ldots, S_{|I|})$ denotes a vector consisting of all basestock levels. The expected number of backorders of LRU i, in steady state (i.e., at an arbitrary point in time in the long run), is denoted by $EBO_i(S_i)$. The aggregate expected number of backorders,

in steady state, is:

$$EBO(\mathbf{S}) = \sum_{i \in I} EBO_i(S_i). \tag{1}$$

The target level for $EBO(\mathbf{S})$ is given by EBO^{obj} and the solution space is:

$$\mathcal{S} = \{\mathbf{S} = (S_1, \ldots, S_{|I|}) \mid S_i \in \mathbb{N}_0, \forall i \in I\}.$$

Hence, in mathematical terms, our optimization problem, Problem (P), is:

$$\begin{array}{ll} \min & C(\mathbf{S}) \\ \text{subject to} & EBO(\mathbf{S}) \leq EBO^{\text{obj}} \\ & \mathbf{S} \in \mathcal{S}. \end{array} \tag{P}$$

Problem (P) has a linear objective function, a nonlinear constraint, and integral decision variables. It thus is a nonlinear integer programming problem.

The expected backorder position $EBO_i(S_i)$ denotes the number of parts of LRU i that is missing in all machines of the installed base together. A part is said to be missing in case a defective part has not been replaced yet by a ready-for-use part because there is no ready-for-use spare part available. Similarly, $EBO(\mathbf{S})$ denotes the total number of missing parts in all machines together, and thus is a measure for the inconvenience due to insufficient stock of ready-for-use spare parts. The constraint on the aggregate expected number of backorders is closely related to an availability constraint, where the availability $A(\mathbf{S})$ denotes the fraction of machines that is not down due to a missing part, or equivalently, the fraction of time that any given machine is not down due to a missing part. This availability is usually called the supply availability and can be calculated as $\frac{\text{MTBM}}{\text{MTBM+MSD}}$ with MTBM being the mean time between maintenance and MSD being the mean supply delay (see, e.g., [18, p. 38]). See Section 3.6.2 for a detailed description of the relation with the availability constraint. In short, if it hardly occurs that any machine has two or more parts missing, then:

$$A(\mathbf{S}) \approx 1 - \frac{1}{N} EBO(\mathbf{S}),$$

where N denotes the total number of machines. Therefore, setting a maximum level EBO^{obj} for the total expected backorder position is equivalent to setting a minimum level $A^{\text{obj}} = 1 - \frac{1}{N} EBO^{\text{obj}}$ for the availability. Notice that active maintenance time or hands-on-tool time is not influenced by the inventory control policy and is therefore not incorporated in the availability measure that we use.

3.2. Overview of assumptions

We summarize and discuss the main assumptions made in Section 3.1:

1. *Demands for the different LRUs occur according to independent Poisson processes.*

   The assumption of independent Poisson processes is justified when a failure of a component does not lead to additional failures of other components in the same machine. In general this is true. The assumption of Poisson processes is justified either when lifetimes of components are exponential or when lifetimes are generally distributed and the number of machines that is served by the warehouse is sufficiently large.
2. *For each LRU, the demand rate is constant.*

   The single warehouse serves multiple machines. When one or more machines fail and defective parts cannot be provided, then some machines are down for a while and the demand rate for a given LRU decreases accordingly. However, when the fraction of machines that is down is always sufficiently small, either because downtimes are short in general or because downtimes occur only rarely, then the decrease in demand rate is small, and thus it is reasonable to assume a constant demand rate. In other cases, workload of a defective machine is routed to another machine, so that the total workload and thus the total failure rate remains the same (until the number of operating machines becomes too small).
3. *Repair leadtimes for different LRUs are independent and repair leadtimes for parts of the same LRU are independent and identically distributed.*

   For repairable LRUs, this assumption is justified if planned repair leadtimes have been agreed upon with repair shops (external companies or departments within the same company). It is then the responsibility of the repair shop to meet the planned leadtimes. In practice, planned leadtimes often occur either because repair is executed by an external company or in order to decompose the inventory control from the control of the repair facilities. An analogous reasoning holds for consumable LRUs.
4. *A one-for-one replenishment strategy is applied for all LRUs.*

   This is justified as long as there are no fixed ordering costs or if the fixed ordering costs are small relative to the prices of the LRUs (or, thinking of the Economic Order Quantity rule, relative to the price divided by the demand rate). If fixed ordering costs are relevant, then fixed order quantities may be appropriate to assume and an (s, Q) policy can be used instead of a basestock policy for each LRU. This extension is described in Section 6.3.

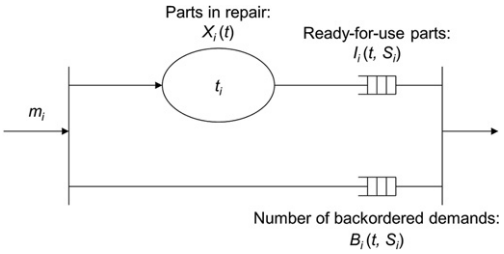

Fig. 3. Petri net of the repair and demand fulfillment process of LRU i.

3.3. Evaluation

In this section, we evaluate the steady-state behavior and the aggregate expected number of backorders $EBO(\mathbf{S})$ for a given basestock policy $\mathbf{S}$. Because parts of different LRUs have no interaction, the steady-state behavior can be evaluated per LRU. This leads to a closed-form expression for $EBO_i(S_i)$. $EBO(\mathbf{S})$ itself then follows from (1). The derivation of the expression for $EBO_i(S_i)$ can be found in [49], although they focus on what they call the expected number of fills and "the expected number of backorders is equal to the expected number of demands minus the expected number of fills" [49, p. 396]. The expression for $EBO_i(S_i)$ was first given explicitly by Sherbrooke [16, p. 132].

Consider an arbitrary LRU i, and assume that the basestock level S_i is given. The repair and demand fulfillment process of this LRU is depicted by the Petri net in Fig. 3. On the left-hand side in this figure, demands for ready-for-use parts, accompanied with defective parts, arrive with rate m_i. The defective parts follow the upper stream in the figure. That is, they first go into repair which takes on average t_i time units. Then they arrive in a queue with ready-for-use parts. Actually this queue represents the physical stock, also called stock on hand. The demands for ready-for-use parts follow the lower stream. That is, these requests are sent to the warehouse, where they are fulfilled immediately if there is enough stock on hand and after some delay otherwise. Delayed requests are fulfilled according to a first-come, first-served (FCFS) discipline. When both a request and a ready-for-use part are available, they merge (i.e., the transition on the right-hand side in the figure 'fires')

and leave the system. It is easily seen that at any point in time at least one of the two queues on the right-hand side in the figure is empty: if the stock on hand is positive, then there will be no requests waiting for a ready-for-use part; if the number of requests in the queue (the number of backorders) is positive, then there will be no part in the queue with on hand stock.

We describe the state of the system at time instant t by $(X_i(t), I_i(t, S_i), B_i(t, S_i))$, where $X_i(t)$ denotes the number of parts in repair at time t, $I_i(t, S_i)$ denotes the stock on hand of ready-for-use parts at time t, and $B_i(t, S_i)$ denotes the number of backordered demands at time t. Both $I_i(t, S_i)$ and $B_i(t, S_i)$ depend on S_i. Notice that $X_i(t)$ does not depend on S_i; it depends only on the Poisson arrival process of defective parts, as seen in Fig. 3. The amount $X_i(t)$ represents the number of parts in the repair pipeline and is therefore also called the *pipeline stock*. Notice that $(X_i(t), I_i(t, S_i), B_i(t, S_i))$ constitutes a partial description only; for a full description, since repair leadtimes are generally distributed, it is also required to denote how long parts have been in repair.

The possible values for the tuples $(X_i(t), I_i(t, S_i), B_i(t, S_i))$ are given by:

$$(0, S_i, 0), (1, S_i - 1, 0), \ldots, (S_i - 1, 1, 0), (S_i, 0, 0),$$
$$(S_i + 1, 0, 1), (S_i + 2, 0, 2), \ldots.$$

The first S_i states in this sequence are with positive stock on hand, the state $(S_i, 0, 0)$ is the unique state where both the stock on hand and the number of backordered demands is zero, and after that the states with a positive number of backordered demands are obtained. A transition is made from one state to the next state in this sequence when a demand occurs, while a completion of a repair leads to a transition from one state to a previous state in this sequence. From the sequence with all possible states, we observe that the values of $I_i(t, S_i)$ and $B_i(t, S_i)$ follow directly from the values of $X_i(t)$ and S_i. It holds that:

$$I_i(t, S_i) = (S_i - X_i(t))^+, \tag{2}$$

$$B_i(t, S_i) = (X_i(t) - S_i)^+, \tag{3}$$

where $x^+ = \max\{0, x\}$ for any $x \in \mathbb{R}$. These equations imply that:

$$I_i(t, S_i) - B_i(t, S_i) = S_i - X_i(t),$$

or, equivalently, that:

$$X_i(t) + I_i(t, S_i) - B_i(t, S_i) = S_i.$$

The latter equation is known as the *stock balance equation* (see, e.g., [18, p. 24]) and shows that the number of parts in the upper stream of the Petri net in Fig. 3 is always S_i more than the number of requests in the lower stream.

Let X_i, $I_i(S_i)$, and $B_i(S_i)$ be the steady-state variables corresponding to $X_i(t)$, $I_i(t, S_i)$, and $B_i(t, S_i)$, respectively. In other words, they are random variables denoting the number of parts in repair, the number of ready-for-use parts, and the number of backordered demands in steady state. By (2) and (3):

$$I_i(S_i) = (S_i - X_i)^+, \tag{4}$$

$$B_i(S_i) = (X_i - S_i)^+. \tag{5}$$

In our model, defective parts enter the repair pipeline according to a Poisson process and each defective part stays on average a time t_i in the repair pipeline. The repair pipeline can be seen as a queueing system with infinitely many servers and service times t_i. In other words, the repair pipeline is an $M/G/\infty$ queueing system and we may thus apply Palm's Theorem [50]:

> *Palm's Theorem*: If jobs arrive according to a Poisson process with rate λ at a service system and if the times that the jobs remain in the service system are independent and identically distributed according to a given general distribution with mean EW, then the steady-state distribution for the total number of jobs in the service system is Poisson distributed with mean λEW.

Application of this theorem to the repair pipeline leads to part (i) of the following lemma; the parts (ii) and (iii) of this lemma follow from part (i) and the Eqs. (4) and (5).

Lemma 3.1. *Let $i \in I$.*

(i) *The pipeline X_i is Poisson distributed with mean $m_i t_i$, i.e.:*

$$\mathbb{P}\{X_i = x\} = \frac{(m_i t_i)^x}{x!} e^{-m_i t_i}, \quad \forall x \in \mathbb{N}_0.$$

(ii) *The distribution of the stock on hand $I_i(S_i)$ is given by:*

$$\mathbb{P}\{I_i(S_i) = x\} = \begin{cases} \sum_{y=S_i}^{\infty} \mathbb{P}\{X_i = y\} & \text{if } x = 0; \\ \mathbb{P}\{X_i = S_i - x\} & \text{if } x \in \{1, \ldots, S_i\}. \end{cases}$$

(iii) *The distribution of the number of backordered demands $B_i(S_i)$ is given by:*

$$\mathbb{P}\{B_i(S_i) = x\} = \begin{cases} \sum_{y=0}^{S_i} \mathbb{P}\{X_i = y\} & \text{if } x = 0; \\ \mathbb{P}\{X_i = x + S_i\} & \text{if } x \in \mathbb{N}. \end{cases}$$

Lemma 3.1 contains the main results for the evaluation of a given policy. From this lemma, we easily obtain relevant service measures, among which the expected backorder positions $EBO_i(S_i)$:

$$\begin{aligned} EBO_i(S_i) &= \mathbb{E}B_i(S_i) = \sum_{x=S_i+1}^{\infty} (x - S_i)\mathbb{P}\{X_i = x\} \\ &= m_i t_i - S_i + \sum_{x=0}^{S_i} (S_i - x)\mathbb{P}\{X_i = x\}, \quad \forall S_i \in \mathbb{N}_0. \end{aligned} \tag{6}$$

Notice that the latter expression for $EBO_i(S_i)$ is most appropriate for computational purposes as it avoids complications because of sums with infinitely many terms. Computations can further be simplified using the fact that for $S_i \in \mathbb{N}$:

$$EBO_i(S_i) = EBO_i(S_i - 1) - 1 + \sum_{x=0}^{S_i - 1} \mathbb{P}\{X_i = x\}.$$

3.4. Optimization

Instead of solving Problem (P) directly, we consider a closely related problem, Problem (Q), with two objectives, minimization of the investment $C(\mathbf{S})$ and minimization of the aggregate expected number of backorders $EBO(\mathbf{S})$:

$$\begin{aligned} &\min && C(\mathbf{S}) \\ &\min && EBO(\mathbf{S}) \\ &\text{subject to} && \mathbf{S} \in \mathcal{S}. \end{aligned} \tag{Q}$$

This problem is a multi-objective programming problem. For this problem, we will derive *efficient solutions*. A solution $\mathbf{S} \in \mathcal{S}$ is efficient for Problem (Q) if and only if there is no other solution $\mathbf{S}' \in \mathcal{S}$ with $C(\mathbf{S}') \leq C(\mathbf{S})$ and $EBO(\mathbf{S}') \leq EBO(\mathbf{S})$, and strict inequality for at least one of these inequalities. Alternatively stated, a solution $\mathbf{S} \in \mathcal{S}$ is efficient for Problem (Q) if and only if $C(\mathbf{S}') > C(\mathbf{S})$, or $EBO(\mathbf{S}') > EBO(\mathbf{S})$, or $(C(\mathbf{S}'), EBO(\mathbf{S}')) = (C(\mathbf{S}), EBO(\mathbf{S}))$ for all $\mathbf{S}' \in \mathcal{S}$. Let $\mathcal{E}^*$ denote the set of all efficient solutions for Problem (Q). Then the points $(C(\mathbf{S}), EBO(\mathbf{S}))$, $\mathbf{S} \in \mathcal{E}^*$, constitute an *efficient frontier* for the total inventory investment versus aggregate expected number of backorders. From this efficient frontier, an optimal solution for Problem (P) can be picked.

The idea of focusing on Problem (Q) instead of solving Problem (P) directly comes from [16], but has first been formalized, to the best of our knowledge, by Van Houtum and Hoen [51].

Problem (Q) has the following structure: $C(\mathbf{S}) = \sum_{i\in I} C_i(S_i)$, $EBO(\mathbf{S}) = \sum_{i\in I} EBO_i(S_i)$, and $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_{|I|}$, where $\mathcal{S}_i = \mathbb{N}_0$ represents the solution space for S_i for all $i \in I$, i.e., the objective functions are separable and the solutions space is a Cartesian Product, and thus Problem (Q) as a whole is *separable* [52]. In addition, for all LRUs $i \in I$ the functions $C_i(S_i) = c_i^{\mathrm{h}} S_i$ are linear, and, as we shall derive in Section 3.4.1, the functions $EBO_i(S_i)$ are decreasing and convex. This has first been shown by Sherbrooke [16, p. 132]. As a result, a greedy procedure can be applied to generate efficient solutions, see Section 3.4.2. This greedy procedure has been proposed by Sherbrooke [16]. He also mentions that convexity is sufficient to guarantee that the greedy procedure finds optimal solutions. A formal proof of this, specifically for our model, can be found in [51].

3.4.1. Convexity of the expected backorder positions

Definition 3.1. Let $f(x)$ be a function on $\mathbb{N}_0$ and $x_0 \in \mathbb{N}_0$.

(i) $f(x)$ is decreasing for $x \geq x_0$ if:

$$\Delta f(x) = f(x+1) - f(x) \leq 0, \quad \forall x \geq x_0.$$

(ii) $f(x)$ is convex for $x \geq x_0$ if:

$$\Delta^2 f(x) = \Delta f(x+1) - \Delta f(x) \geq 0, \quad \forall x \geq x_0.$$

Notice that $\Delta f(x+1) - \Delta f(x) = f(x+2) - 2f(x+1) + f(x)$, $\forall x \in \mathbb{N}_0$. The definitions for strictly decreasing and strictly convex are obtained by replacing the inequality signs by strict inequality signs. The definitions for (strictly) increasing and (strictly) concave are obtained by turning the (strict) inequality signs around.

The expected number of backorders $EBO_i(S_i)$ for LRU $i \in I$ is a function on $\mathbb{N}_0$. Lemma 3.2 states that $EBO_i(S_i)$ is decreasing and convex on its whole domain.

Lemma 3.2. *For each LRU $i \in I$, $EBO_i(S_i)$ is decreasing and convex for $S_i \in \mathbb{N}_0$.*

Proof. Let $i \in I$. By (6):

$$\begin{aligned}\Delta EBO_i(S_i) &= EBO_i(S_i+1) - EBO_i(S_i) \\ &= -\sum_{x=S_i+1}^{\infty} \mathbb{P}\{X_i = x\} \leq 0, \quad \forall S_i \in \mathbb{N}_0, \end{aligned} \tag{7}$$

which shows that $EBO_i(S_i)$ is decreasing on its whole domain. Further:

$$\begin{aligned}\Delta^2 EBO_i(S_i) &= \Delta EBO_i(S_i+1) - \Delta EBO_i(S_i) \\ &= \mathbb{P}\{X_i = S_i + 1\} \geq 0, \quad \forall S_i \in \mathbb{N}_0,\end{aligned}$$

which shows that $EBO_i(S_i)$ is convex on its whole domain. □

3.4.2. Greedy algorithm

Problem (Q) is separable and the functions $EBO_i(S_i)$ are decreasing and convex on their whole domains. Hence we can prove that a set of efficient solutions can be generated by a greedy algorithm. A first efficient solution $\mathbf{S} = (S_1, \ldots, S_{|I|})$ is obtained by setting $S_i = 0$ for each LRU $i \in I$. This solution is efficient because it has the lowest possible investment $C(\mathbf{S}) = 0$. Next, for each LRU i, we compute the decrease in $EBO(\mathbf{S})$ relative to the increase in $C(\mathbf{S})$ when S_i would be increased by one unit. The increase in $C(\mathbf{S})$ equals c_i^{h}, while the change in $EBO(\mathbf{S})$ equals (using (7)):

$$\begin{aligned}\Delta_i EBO(\mathbf{S}) &= \Delta EBO_i(S_i) = -\sum_{x=S_i+1}^{\infty} \mathbb{P}\{X_i = x\} \\ &= -\left(1 - \sum_{x=0}^{S_i} \mathbb{P}\{X_i = x\}\right).\end{aligned}$$

The decrease in $EBO(\mathbf{S})$, which is equal to $-\Delta_i EBO(\mathbf{S})$, divided by the increase in $C(\mathbf{S})$, equal to c_i^{h}, is denoted by Γ_i. The LRU with the highest value for Γ_i is selected (also referred to as 'biggest bang for the buck'), and the corresponding basestock level is increased by one unit (ties may be broken arbitrarily). The new solution $\mathbf{S}$ is also efficient and is added to a set of efficient solutions. The generation of efficient solutions is continued until a given aggregate expected number of backorders or inventory investment has been reached. The formal procedure is described in Algorithm 3.1, where e_k is an I-dimensional unit row-vector consisting of $|I| - 1$ zeros and a one at the kth position.

Algorithm 3.1 (*Greedy Algorithm*)**.**

Step 1. $S_i := 0$ for all $i \in I$ (so $\mathbf{S} := (0, \ldots, 0)$);
$\mathcal{E} := \{\mathbf{S}\}$;
$C(\mathbf{S}) := 0$ and $EBO(\mathbf{S}) := \sum_{i\in I} m_i t_i$.

Step 2. $\Gamma_i := \frac{1}{c_i^{\mathrm{h}}}\left(1 - \sum_{x=0}^{S_i} \mathbb{P}\{X_i = x\}\right)$ for all $i \in I$;
$k := \arg\max_{i\in I} \Gamma_i$;
$\mathbf{S} := \mathbf{S} + e_k$;
$\mathcal{E} := \mathcal{E} \cup \{\mathbf{S}\}$.

Step 3. $C(\mathbf{S}) := C(\mathbf{S}) + c_k^{\mathrm{h}}$;
$EBO(\mathbf{S}) := EBO(\mathbf{S}) - 1 + \sum_{x=0}^{S_k} \mathbb{P}\{X_k = x\}$;
If 'stop criterium', then stop, else go to step 2.

In the following lemma, it is formally stated that Algorithm 3.1 generates efficient solutions for Problem (Q). The proof of this lemma follows directly from Theorem 2 in [52].

Lemma 3.3. *At termination of Algorithm 3.1, the set $\mathcal{E}$ consists of efficient solutions for Problem (Q), i.e., $\mathcal{E} \subset \mathcal{E}^*$.*

For the computation of the pipeline stock probabilities $\mathbb{P}\{X_i = x\}$ in Algorithm 3.1, we advice the use of the following recursion for the sake of efficiency and to avoid numerical problems:

$$\mathbb{P}\{X_i = x\} = \begin{cases} e^{-m_i t_i} & \text{if } x = 0; \\ \dfrac{m_i t_i}{x}\mathbb{P}\{X_i = x-1\} & \text{if } x \in \mathbb{N}. \end{cases} \tag{8}$$

In general, the greedy algorithm generates an ordered set $\mathcal{E} = \{\mathbf{S}^0, \mathbf{S}^1, \mathbf{S}^2, \ldots\}$ of efficient solutions for Problem (Q), where $EBO(\mathbf{S}^0) > EBO(\mathbf{S}^1) > EBO(\mathbf{S}^2) > \cdots$ and $0 = C(\mathbf{S}^0) < C(\mathbf{S}^1) < C(\mathbf{S}^2) < \cdots$. For Problem (P) with a given target EBO^{obj}, obtaining a feasible solution from the set $\mathcal{E}$ generated by the greedy algorithm is straightforward: take the first solution $\mathbf{S}^l \in \mathcal{E}$ with $EBO(\mathbf{S}^l) \leq EBO^{\mathrm{obj}}$. This solution is optimal if and only if there is no solution $\mathbf{S} \in \mathcal{E}^*$ with $EBO(\mathbf{S}^l) < EBO(\mathbf{S}) \leq EBO^{\mathrm{obj}}$. In general, the solution $\mathbf{S}^l$ will be close to optimal if $EBO(\mathbf{S}^l)$ is close to EBO^{obj}. The optimality gap will be larger if at the last iteration of the algorithm, the basestock level of an LRU is increased that gives a large jump for the aggregate expected number of backorders (and the corresponding costs). For real-life problems, such large jumps are not very likely because there are many LRUs.

We can conclude that for instances with sufficiently many LRUs, the greedy algorithms will generate good heuristic solutions for Problem (P). Besides, a greedy algorithm is efficient, it is easy to implement in practice, and it is an algorithm that is easy to understand by practitioners. To find an optimal solutions, Problem (P) can be solved by a similar algorithm as for knapsack problems. A disadvantage of those algorithms is that a small change in input parameters (cost prices of the LRUs, demand rates, or the target EBO^{obj}) may lead to large changes in the optimal solution. The solution generated by the greedy algorithm, however, will be rather robust.

3.5. Alternative optimization techniques

We can also solve Problem (P) with other techniques rather than the greedy algorithm of Section 3.4.2: we describe Lagrange relaxation in Section 3.5.1 and we briefly mention Dantzig–Wolfe decomposition in Section 3.5.2.

3.5.1. Lagrange relaxation

We apply the Lagrangian relaxation technique to Problem (P); for a general description of this technique, we refer to Appendix B of [53]. The *Lagrangian* for (P) is defined as:

$$L(\mathbf{S},\lambda)=\sum_{i\in I}c_i^{\text{h}}S_i+\lambda\left(\sum_{i\in I}EBO_i(S_i)-EBO^{\text{obj}}\right),$$

where $\lambda\geq 0$ is a *Lagrange multiplier*.

It has been noticed before that Problem (Q) is separable. This also holds for Problem (P) (see also the definition of separable problems in [53, Appendix B]). We can separate Problem (P) because it is a linear combination of LRU objectives and constraints. It is known that in separable problems, the Lagrangian is also separable. The Lagrangian can be rewritten as:

$$L(\mathbf{S},\lambda)=\sum_{i\in I}L_i(S_i,\lambda)-\lambda EBO^{\text{obj}}, \tag{9}$$

where:

$$L_i(S_i,\lambda)=c_i^{\text{h}}S_i+\lambda EBO_i(S_i)$$

is the *decentralized Lagrangian* for LRU i. Notice that, in Eq. (9), we have $|I|$ different Lagrangians, one for every LRU. Notice also that we have only one λ because we have only one constraint in our problem.

For any given value of λ, we can find a base stock level that minimizes the decentralized Lagrangian (for every LRU i). Since the decentralized Lagrangians are convex functions, we know that it has either one unique minimum or multiple minima in subsequent points. One way to find this minimum is to start with $S_i=0$ and increase it by 1 at a time, until the values for the decentralized Lagrangian start increasing. The resulting base stock vector is a solution to problem (P). We can now vary the value of λ to find different solutions to Problem (P). Then, we calculate the corresponding value of $EBO(\mathbf{S})$ and $C(\mathbf{S})$. As soon as we find a value of $EBO(\mathbf{S})$ that is at or below our target value, we can stop the procedure.

Using the Lagrange relaxation method gives us optimal solutions of Problem (P) for specific values of EBO^{obj}. This follows from the Everett Result [54], which for our problem reads as follows:

> *The Everett result*: If, for a given $\lambda\geq 0$, $\mathbf{S}(\lambda)$ minimizes $L(\mathbf{S},\lambda)$ over $\mathbf{S}\in\mathcal{S}$, then $\mathbf{S}(\lambda)$ is optimal for Problem (P) for every $EBO^{\text{obj}}\in(0,\infty)$ that satisfies
>
> $$EBO^{\text{obj}}\geq EBO(\mathbf{S}(\lambda)) \quad\text{and}\quad \lambda\left(EBO(\mathbf{S}(\lambda))-EBO^{\text{obj}}\right)=0.$$

If we take $\lambda=0$, then each Lagrangian $L_i(S_i,\lambda)$ is strictly increasing, and we find $\mathbf{S}(0)=(0,\ldots,0)$ and $EBO(\mathbf{S}(0))=\sum_{i\in I}m_it_i$. The solution $\mathbf{S}(0)=(0,\ldots,0)$ is optimal for Problem (P) for every $EBO^{\text{obj}}\geq\sum_{i\in I}m_it_i$. For each $\lambda>0$, the solution $\mathbf{S}(\lambda)$ is optimal for Problem (P) for $EBO^{\text{obj}}=EBO(\mathbf{S}(\lambda))$. I.e., then the optimality of $\mathbf{S}(\lambda)$ is guaranteed for one specific value of EBO^{obj} (but the solution might also be optimal for slightly higher values of EBO^{obj}).

The Lagrange relaxation method gives efficient solutions for Problem (Q). This follows directly from Theorem 1 in [52]. In fact, the Lagrange relaxation yields exactly the same solutions as the Greedy approach. This is not a coincidence. When we study the details of the execution of the greedy procedure and the execution of the Lagrangian relaxation method, we see the similarities. The key is a one-to-one relationship between the Γ_i values computed in the greedy algorithm and the values of λ for which $\mathbf{S}(\lambda)$ changes to the next solution in the Lagrange relaxation method.

3.5.2. Dantzig–Wolfe decomposition

Dantzig–Wolfe decomposition, as introduced by Dantzig and Wolfe [55], can also be applied to Problem (P). For details of this method applied to our problem, see Section 1.4.2 of [1]. It appears that this method yields exactly the same solutions as the Lagrange relaxation method; for further references about this parallel, see [1]. Both methods can be used more or less interchangeably and we will see examples of both methods applied to more complicated models in the remainder of this paper. An advantage of using Dantzig–Wolfe decomposition is that it facilitates the usage of standard mathematical programming techniques (this is especially relevant for problems with multiple constraints, like the ones that we discuss in Section 4).

3.6. Alternative service measures

Now consider the situation that we are interested in some other service measure instead of the aggregate expected number of backorders. Fortunately, it is possible to adjust the greedy algorithm procedure to incorporate a different service constraint. There are many alternative service measures possible; we discuss the adjustments required for three of those that are used relatively frequently. In Section 3.6.1, we discuss the implications of using the expected waiting time. We do the same for the availability in Section 3.6.2 and for the aggregate fill rate in Section 3.6.3.

3.6.1. Expected waiting time

It is straightforward to adapt the service level constraint based on the expected number of backorders to one that is based on the expected waiting time until an arbitrary spare parts demand is fulfilled. The expected waiting time for a spare part of LRU $i\in I$ when the base stock level is S_i, can be determined using Little's formula [56]: $W_i(S_i)=EBO_i(S_i)/m_i$. Taking all LRUs together, the aggregate expected waiting time $W(\mathbf{S})$ is:

$$\begin{aligned}W(\mathbf{S}) &= \sum_{i\in I}\mathbb{P}\{\text{an arbitrary demand is for LRU } i\}\\ &\quad\times(\text{expected waiting time for LRU } i)\\ &= \sum_{i\in I}\frac{m_i}{M}\times\frac{EBO_i(S_i)}{m_i}=\frac{1}{M}\sum_{i\in I}EBO_i(S_i).\end{aligned}$$

This means that we get a problem that is equivalent to the original problem.

3.6.2. Average availability

As stated in Section 3.1, and by Sherbrooke [18], the constraint on the aggregate expected number of backorders is closely related to an availability constraint. The average availability is equal to

the fraction of time that any given machine is available. Let N be the total number of machines, and let Z_i be the number of parts of LRU i installed per machine. We can approximate the average availability as follows. The average number of backorders of LRU i is $EBO_i(S_i)$. Hence, the probability that a given part of LRU i in a given machine is working is equal to $1 - \frac{EBO_i(S_i)}{NZ_i}$. Next, ignoring dependencies between these probabilities for the various parts in a given machine, we obtain the following approximation for $A(\mathbf{S})$:

$$A(\mathbf{S}) \approx \prod_{i\in I}\left(1 - \frac{EBO_i(S_i)}{NZ_i}\right)^{Z_i}.$$

For sufficiently high values of $A(\mathbf{S})$, the product on the right hand side may be approximated by its first order approximation:

$$\begin{aligned} A(\mathbf{S}) &\approx 1 - \sum_{i\in I} Z_i \frac{EBO_i(S_i)}{NZ_i} = 1 - \frac{1}{N}\sum_{i\in I} EBO_i(S_i) \\ &= 1 - \frac{1}{N} EBO(\mathbf{S}). \end{aligned}$$

Hence, for a sufficiently high target A^{obj} for $A(\mathbf{S})$, a heuristic solution for the problem with a target average availability can be obtained via the heuristic solution for Problem (P) with target $EBO^{\text{obj}} = N(1 - A^{\text{obj}})$ for the aggregate expected number of backorders.

3.6.3. Aggregate fill rate

The aggregate fill rate is defined as the probability that an arbitrary demand for the total group of LRUs is fulfilled immediately, or, equivalently, as the fraction of the total demand stream that is fulfilled from stock. Let the fill rate for LRU i, also called item fill rate, be denoted by $\beta_i(S_i)$, then:

$$\beta(\mathbf{S}) = \sum_{i\in I} \frac{m_i}{M}\beta_i(S_i). \tag{10}$$

The target aggregate fill rate is given by β^{obj}. This service measure has already been used for spare parts in the 1960s (see, e.g., [57]).

Demands for LRU i arrive according to a Poisson process, and thus, by the PASTA (Poisson Arrivals See Time Averages) property, an arbitrary arriving demand observes the system in steady state. Hence, with probability $\mathbb{P}\{I_i(S_i) > 0\} = \mathbb{P}\{X_i < S_i\}$, a positive stock on hand is observed and the demand can be fulfilled immediately, and otherwise not. Therefore:

$$\beta_i(S_i) = \sum_{x=0}^{S_i-1} \mathbb{P}\{X_i = x\}. \tag{11}$$

The item fill rate $\beta_i(S_i)$ for an LRU $i \in I$ is a function on $\mathbb{N}_0$. Lemma 3.4 states that $\beta_i(S_i)$, and thus also $f_i(S_i) = \frac{m_i}{M}\beta_i(S_i)$, is increasing on its whole domain and concave for $S_i \geq \max\{\lceil m_i t_i - 1\rceil, 0\}$, where $\lceil x\rceil$ denotes rounding up x to the next higher integer (i.e., $\lceil x\rceil = x'$ for any $x \in \mathbb{R}$, $x' \in \mathbb{N}_0$ and $x \leq x' < x + 1$).

Lemma 3.4. *For each LRU $i \in I$, the item fill rate $\beta_i(S_i)$ is increasing on its whole domain and concave for $S_i \geq \max\{\lceil m_i t_i - 1\rceil, 0\}$.*

Proof. Let $i \in I$. By (11):

$$\begin{aligned} \Delta\beta_i(S_i) &= \beta_i(S_i + 1) - \beta_i(S_i) \\ &= \mathbb{P}\{X_i = S_i\} \geq 0, \quad \forall S_i \in \mathbb{N}_0, \end{aligned} \tag{12}$$

which shows that $\beta_i(S_i)$ is increasing on its whole domain. Further:

$$\Delta^2\beta_i(S_i) = \mathbb{P}\{X_i = S_i + 1\} - \mathbb{P}\{X_i = S_i\}, \quad \forall S_i \in \mathbb{N}_0. \tag{13}$$

By (8):

$$\mathbb{P}\{X_i = S_i + 1\} = \frac{m_i t_i}{S_i + 1}\mathbb{P}\{X_i = S_i\}, \quad \forall S_i \in \mathbb{N}_0,$$

and by substitution of this recursive relation into (13), we find:

$$\Delta^2\beta_i(S_i) = \left(\frac{m_i t_i}{S_i + 1} - 1\right)\mathbb{P}\{X_i = S_i\}, \quad \forall S_i \in \mathbb{N}_0.$$

From this formula, it follows that $\Delta^2\beta_i(S_i) \leq 0$ if and only if $\frac{m_i t_i}{S_i+1} - 1 \leq 0$, i.e., if and only if $S_i \geq m_i t_i - 1$. In other words, $\beta_i(S_i)$ is concave for $S_i \geq m_i t_i - 1$. Because of the integrality and non-negativity of S_i, the condition $S_i \geq m_i t_i - 1$ is equivalent to $S_i \geq \max\{\lceil m_i t_i - 1\rceil, 0\}$. □

The amount $m_i t_i$ represents the average number of parts in the repair pipeline. If this average pipeline stock is smaller than or equal to 1, then $\max\{\lceil m_i t_i - 1\rceil, 0\} = 0$ and thus $\beta_i(S_i)$ is concave on its whole domain. If the average pipeline stock is larger than 1, then $\max\{\lceil m_i t_i - 1\rceil, 0\} > 0$ and $\beta_i(S_i)$ is not concave on its whole domain but, roughly spoken, on the right-hand side of the average pipeline stock, which is the relevant part for optimization purposes.

We now reformulate Problem (Q) as defined in Section 3.4. First, we replace the minimization of $EBO(\mathbf{S})$ by the maximization of $\beta(\mathbf{S})$. Second, we limit the solution space to $\mathcal{S}' = \mathcal{S}'_1 \times \mathcal{S}'_2 \times \cdots \times \mathcal{S}'_{|I|}$, with $\mathcal{S}'_i = \{S_i \in \mathbb{N}_0 \mid S_i \geq m_i t_i - 1\}$ for all $i \in I$. Hence, we obtain the Problem (Q′):

$$\begin{aligned} \min \quad & C(\mathbf{S}) \\ \max \quad & \beta(\mathbf{S}) \\ \text{subject to} \quad & \mathbf{S} \in \mathcal{S}'. \end{aligned} \tag{Q′}$$

In this problem, solutions with small S_i are excluded. Obviously, such solutions are not relevant for high target aggregate fill rates.

Problem (Q′) is still separable and the functions $f_i(S_i)$ are now increasing and concave on their whole domains $\mathcal{S}'$. Hence a set of efficient solutions can be generated by a greedy algorithm. A first efficient solution $\mathbf{S} = (S_1, \ldots, S_{|I|})$ is obtained by setting $S_i = \max\{\lceil m_i t_i - 1\rceil, 0\}$ for each LRU $i \in I$. This solution is efficient because it has the lowest possible investment. Next, we execute greedy steps. In each step, we compute for each LRU i the increase in $\beta(\mathbf{S})$ relative to the increase in $C(\mathbf{S})$ when S_i would be increased by one unit. The increase in $C(\mathbf{S})$ equals c_i^{h}, while the increase in $\beta(\mathbf{S})$ equals (using (12)):

$$\Delta_i\beta(\mathbf{S}) = \Delta f_i(S_i) = \frac{m_i}{M}\left(\beta_i(S_i + 1) - \beta(S_i)\right) = \frac{m_i}{M}\mathbb{P}\{X_i = S_i\}.$$

The increase in $\beta(\mathbf{S})$ divided by the increase in $C(\mathbf{S})$ is denoted by Γ_i. The LRU with the highest value for Γ_i is selected, and the corresponding basestock level is increased.

4. METRIC model

In Section 3, we have discussed single-location models. In practice, however, spare parts networks usually consist of multiple echelon levels, see Section 2. In the current section, we discuss the simplest example of such a network, a two-echelon network, consisting of a number of local warehouses and a central depot. The model that we discuss has been proposed in the seminal work of Sherbrooke [16], except that he uses a different service constraint: Sherbrooke uses an aggregate service level target over all local warehouses, whereas we use a service level target per local warehouse. The reason is that we believe that the latter is more useful in practice. We discuss usage of the former service constraint in Section 4.5.2.

We start with a description of the model in Section 4.1 and an overview of the key assumptions in Section 4.2. There are many similarities with the model and assumptions as discussed in Sections 3.1 and 3.2, respectively. We next discuss evaluation of this model in Section 4.3, and a greedy heuristic for the optimization problem in Section 4.4. We conclude with a discussion of alternative service measures in Section 4.5.

4.1. Model description

We have a non-empty set J^{loc} of local warehouses, numbered $j = 1, \ldots, |J^{\text{loc}}|$. Each local warehouse serves a number of technical systems that are all the same or at least similar. Each technical system consists of a non-empty set of all LRUs, numbered $i = 1, \ldots, |I|$. We assume that the total stream of failures of LRU $i \in I$ as observed by local warehouse $j \in J^{\text{loc}}$ constitutes a Poisson process with a constant rate $m_{i,j}$ (≥ 0). For at least one LRU i and local warehouse j, it holds that $m_{i,j} > 0$. Apart from the local warehouses, there exists a central depot, denoted by index 0. Let J denote the set of all stock points, i.e., $J = \{0\} \cup J^{\text{loc}}$.

If a part of LRU i fails at a local warehouse j, a spare part stocked at the local warehouse j is used to replace the defective part, if possible. Otherwise, a backorder arises, until a spare part becomes available. Upon failure, also immediately a replenishment order is placed at the central depot. Define $m_{i,0} := \sum_{j \in J^{\text{loc}}} m_{i,j}$ as the total demand rate for LRU i at the central depot. The demand at the central depot is also a Poisson process since it is the superposition of the Poisson demand processes at the local warehouses. The replenishment order arrives after a *deterministic* leadtime $t_{i,j}$ (order-and-ship time), if stock is available at the central depot. Otherwise, the order is backordered until a spare part becomes available at the central depot. The defective part at the local warehouse is immediately sent to the central depot to be repaired there. It takes a certain random leadtime with mean $t_{i,0}$ before the defective part is repaired and back in stock at the central depot. Equivalently, from a modeling point of view, the defective part can be scrapped and after a certain random leadtime, a newly purchased part is back in stock at the central depot. As in Section 3, we will use the terminology of repairable parts. Notice that we thus assume that each LRU i at each stock point j is controlled according to a base stock policy, with base stock level $S_{i,j}$. The policy in the total network is defined by the $|I| \times |J|$ matrix $\mathbf{S}$, consisting of elements $S_{i,j}$. Each column in this matrix, denoted by a vector $\mathbf{S}_j$, consists of all base stock levels at stock point $j \in J$.

We assume that, for each LRU $i \in I$, backordered replenishment orders from the local warehouses at the central depot are fulfilled in first-come, first-served (FCFS) order. A holding cost c_i^{h} is counted per spare part of LRU i, and the aggregate holding costs are given by:

$$C(\mathbf{S}) = \sum_{i \in I} \sum_{j \in J} c_i^{\text{h}} S_{i,j}.$$

The expected number of backorders for an LRU $i \in I$ at local warehouse $j \in J^{\text{loc}}$ at an arbitrary point in time at the long run, is given by $EBO_{i,j}(S_{i,0}, S_{i,j})$ (notice that this number only depends on $S_{i,0}$ and $S_{i,j}$), and the aggregate expected number of backorders is:

$$EBO_j(\mathbf{S}_0, \mathbf{S}_j) = \sum_{i \in I} EBO_{i,j}(S_{i,0}, S_{i,j}).$$

At local warehouse j, there is a maximum level EBO_j^{obj} given for the aggregate expected number of backorders (a target for all warehouses together is also possible, see Section 4.5.2). Our goal is to determine a system's stocking policy $\mathbf{S}$ to minimize the total holding cost subject to a target for the aggregate expected numbers of backorders per local warehouse.

Our optimization problem, Problem (R), can be formulated as:

$$\begin{array}{lll} \min & C(\mathbf{S}) & \\ \text{subject to} & EBO_j(\mathbf{S}_0, \mathbf{S}_j) \leq EBO_j^{\text{obj}}, & \forall j \in J^{\text{loc}} \\ & S_{i,j} \in \mathbb{N}_0, & \forall i \in I, \ \forall j \in J. \end{array} \tag{R}$$

Note that in the case of consumable spare parts, it may be more logical to exclude the average number of items in transportation in the calculation of the inventory holding costs; this change is easily made since this part of the costs is constant and equals $\sum_{i \in I} \sum_{j \in J} c_i^{\text{h}} m_{i,j} t_{i,j}$ (by Little's law: [56]).

4.2. Overview of assumptions

In addition to the assumptions that we have discussed in Section 3.2, there are two assumptions that we consider in more detail here:

- *For each LRU, the order-and-ship times are assumed to be deterministic*

  The order-and-ship times are the leadtimes between the central depot and the local warehouses. They consist of administrative delays, order picking, and actual transportation times, which can all be controlled well. This also holds for transportation leadtimes which are specified in contractual agreements with third party logistics providers. As a result, the total leadtime will be quite stable and assuming deterministic order-and-ship times is reasonable. For exact evaluations, the assumption of deterministic order-and-ship times is necessary, while approximate evaluations can be adapted for uncertainty in these times.
- *Replenishment orders at the central depot are fulfilled in FCFS order*

  It intuitively makes a lot of sense to use a FCFS discipline at the central depot, and it is most easy to implement in practice. However, if a local warehouse still has stock on hand and is first in the queue at the central depot, while a second local warehouse has backorders (and zero stock on hand) and is second in the queue at the central depot, it is definitely better to send the first ready-for-use part at the central depot to the second local warehouse. The latter issue is an allocation issue that can be addressed at the operational planning level. The METRIC model itself supports inventory level decisions at the tactical planning level and then a simplifying FCFS service discipline assumption is justified. This simplifying assumption is also supported by the literature on multi-echelon distribution systems: the cost difference under FCFS and optimal allocation is relatively small (see, e.g., [58]).

These two assumptions, together with the assumptions of base stock control and Poisson demand processes with constant rates (as already discussed for the single-location model), are key for obtaining a simple and efficient exact evaluation procedure. Due to these assumptions, all demand processes at the most downstream local warehouses immediately propagate to higher echelons, and simple recursive relations are obtained for pipeline stocks and net stocks by going from upstream to downstream locations.

4.3. Evaluation

For a given policy $\mathbf{S}$, evaluation of the steady-state behavior can be done exactly, as described for the first time by Graves [59]. The proof of the following results is analogous to the proof of Lemma 3.1.

Define $X_{i,0}$ as the total amount on order for LRU $i \in I$ at the central depot in steady state, i.e., the total number of parts that is in repair at the central depot, also called the pipeline. $X_{i,0}$ is Poisson distributed with mean $m_{i,0}t_{i,0}$, i.e.:

$$\mathbb{P}\{X_{i,0} = x\} = \frac{(m_{i,0}t_{i,0})^x}{x!} e^{-m_{i,0}t_{i,0}}, \quad \forall x \in \mathbb{N}_0.$$

Let $I_{i,0}(S_{i,0})$ be the stock on hand for LRU $i \in I$ at the central depot, as a function of the base stock level $S_{i,0}$. Its distribution is given by:

$$\mathbb{P}\{I_{i,0}(S_{i,0}) = x\} = \begin{cases} \sum_{y=S_{i,0}}^{\infty} \mathbb{P}\{X_{i,0} = y\} & \text{if } x = 0; \\ \mathbb{P}\{X_{i,0} = S_{i,0} - x\} & \text{if } x \in \{1, \ldots, S_{i,0}\}. \end{cases}$$

Next, define $B_{i,0}(S_{i,0})$ as the number of backordered demands, for LRU $i \in I$ at the central depot, as a function of $S_{i,0}$. It is distributed as follows:

$$\mathbb{P}\left\{B_{i,0}(S_{i,0}) = x\right\} = \begin{cases} \sum_{y=0}^{S_{i,0}} \mathbb{P}\left\{X_{i,0} = y\right\} & \text{if } x = 0; \\ \mathbb{P}\left\{X_{i,0} = S_{i,0} + x\right\} & \text{if } x \in \mathbb{N}. \end{cases}$$

We now define $B_{i,0}^{(j)}(S_{i,0})$ as the number of backorders of LRU $i \in I$ of local warehouse $j \in J^{\text{loc}}$ in the backorder queue at the central depot. As each backordered demand at the central depot stems from local warehouse j with probability $m_{i,j}/m_{i,0}$, the probability distribution of $B_{i,0}^{(j)}(S_{i,0})$ is obtained by:

$$\mathbb{P}\left\{B_{i,0}^{(j)}(S_{i,0}) = x\right\} = \sum_{y=x}^{\infty} \binom{y}{x} \left(\frac{m_{i,j}}{m_{i,0}}\right)^{x} \left(1 - \frac{m_{i,j}}{m_{i,0}}\right)^{y-x} \mathbb{P}\left\{B_{i,0}(S_{i,0}) = y\right\}. \quad (14)$$

Let, for each LRU $i \in I$ and local warehouses $j \in J^{\text{loc}}$, $Y_{i,j}$ be defined as the total demand during the order-and-ship time $t_{i,j}$, and $X_{i,j}(S_{i,0})$ be defined as the total amount on order given base stock level $S_{i,0}$. It holds that $X_{i,j}(S_{i,0}) = B_{i,0}^{(j)}(S_{i,0}) + Y_{i,j}$. This summation is allowed since the order-and-ship time is deterministic (see, e.g., [19, pp. 70–71], for an extensive explanation). From the distribution of $X_{i,j}(S_{i,0})$, we can derive the distribution of $I_{i,j}(S_{i,0}, S_{i,j})$, the physical stock for LRU i at local warehouse j, as a function of the base stock levels $S_{i,0}$ and $S_{i,j}$. The same holds for the distribution of $B_{i,j}(S_{i,0}, S_{i,j})$, the backorder position for LRU i at local warehouse j. Both derivations are analogous to our derivation for the central depot above. It is now easy to obtain the expected backorder positions $EBO_{i,j}(S_{i,0}, S_{i,j})$, analogous to Eq. (6):

$$EBO_{i,j}(S_{i,0}, S_{i,j}) = \mathbb{E}B_{i,j}(S_{i,0}, S_{i,j}) = \sum_{x=S_{i,j}+1}^{\infty} (x - S_{i,j})\mathbb{P}\{X_{i,j}(S_{i,0}) = x\}$$

$$= m_{i,j}t_{i,j} - S_{i,j} + \sum_{x=0}^{S_{i,j}} (S_{i,j} - x)\mathbb{P}\{X_{i,j}(S_{i,0}) = x\}, \quad \forall S_{i,j} \in \mathbb{N}_0.$$

Exact evaluation as explained above, leads to a computational issue, since calculating the probabilities $\mathbb{P}\left\{X_{i,j}(S_{i,0}) = x\right\}$ for all $i \in I$ and $j \in J$, requires calculating the probabilities $\mathbb{P}\{B_{i,0}(S_{i,0}) = y\}$ for all values $y \in \mathbb{N}_0$ (Eq. (14)). In practice, however, we limit ourselves to $y \in \{0, \ldots, b_i^{\max}\}$, with $b_i^{\max} = \min\{y \mid \mathbb{P}\{B_{i,0} \le y\} \ge 1 - \epsilon\}$ and $\epsilon = 10^{-6}$. We allocate the remaining probability mass $1 - \mathbb{P}\{B_{i,0} \le b_i^{\max}\}$ to $\mathbb{P}\left\{B_{i,0} = b_i^{\max}\right\}$ [1, pp. 154–155].

For large systems with many LRUs and local warehouses, the computational effort of the exact evaluations can become too high, and then approximate evaluation methods like METRIC or Graves' approximation can be used. The METRIC approximation assumes that successive replenishment actions at the local warehouses are independent processes, meaning that the variables $X_{i,j}(S_{i,0})$ are Poisson distributed with mean $(m_{i,j}/m_{i,0})\mathbb{E}B_{i,0}(S_{i,0}) + m_{i,j}t_{i,j}$ (a single-moment fit). Graves [59] proposes an approximate evaluation method based on two-moment fits of negative binomial distributions on the variables $X_{i,j}(S_{i,0})$. This two-moments procedure leads to accurate approximations in all cases, while the METRIC approximation is good in many cases, but it does lead to large deviations in several other cases (especially when $\mathbb{E}B_{i,0}(S_{i,0})$ is large). Wong et al. [60], for instance, present the results of experiments evaluating the accuracy of both approximate evaluation methods when used for executing the greedy procedure (for a target expected waiting time as we discuss in Section 4.5.1).

4.4. Greedy algorithm

A feasible solution can be obtained in an efficient way via a greedy procedure similar to the procedures described in [61,60], with constraints on the expected waiting times. The basic idea of this procedure is to add units of stock in an iterative way, as in Section 3.4. The difference is that at each iteration, we add one unit of stock for an LRU $i \in I$ at a stock point $j \in J$ such that we gain the largest decrease in *distance to the set of feasible solutions* per extra unit of additional cost. The procedure is terminated when a feasible solution is obtained. Note that this feasible solution is not necessarily an efficient solution. Note furthermore that the greedy algorithm that Sherbrooke [16] proposes is slightly different and does guarantee to find an efficient solution. However, that algorithm cannot be used with our service measure.

Let $e_{i,j}$ be a matrix having the same structure as the basestock matrices $\mathbf{S}$, with a one at the position corresponding to LRU $i \in I$ and stockpoint $j \in J$, and with zeros at all other positions. Further, in the description below, we denote the variables $EBO_j(\mathbf{S}_0, \mathbf{S}_j)$ simply as $EBO_j(\mathbf{S})$. The greedy procedure starts by setting all base stock levels equal to zero. We define for each solution $\mathbf{S}$ the distance to the set of feasible solutions as:

$$\sum_{j \in J^{\text{loc}}} \left(EBO_j(\mathbf{S}) - EBO_j^{\text{obj}}\right)^{+}.$$

In each iteration, for each combination of $i \in I$ and $j \in J$, we calculate the reduction of the distance to the set of feasible solutions:

$$\Delta_{i,j}EBO = \sum_{l \in J^{\text{loc}}} \left[\left(EBO_l(\mathbf{S}) - EBO_l^{\text{obj}}\right)^{+} - \left(EBO_l(\mathbf{S} + e_{i,j}) - EBO_l^{\text{obj}}\right)^{+} \right],$$

and we compute the ratio $\Gamma_{i,j} = \Delta_{i,j}EBO/c_i^{\text{h}}$. The basestock level corresponding to the LRU and stock point with the highest value for $\Gamma_{i,j}$ is increased by one unit (ties may be broken arbitrarily). The algorithm stops when a feasible solution is found. A formal description of the greedy procedure is given below as Algorithm 4.1:

Algorithm 4.1 (*Greedy Algorithm*)**.**

Step 1. $S_{i,j} := 0$ for all $i \in I, j \in J$ (so $\mathbf{S}_j := (0, \ldots, 0)$ for all $j \in J$ and $\mathbf{S} := (\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_{|J^{\text{loc}}|})$);
$C(\mathbf{S}) := 0$ and $EBO_j(\mathbf{S}) := \sum_{i \in I} m_{i,j}(t_{i,0} + t_{i,j})$ for all $j \in J^{\text{loc}}$.

Step 2. $\Gamma_{i,j} := \frac{\Delta_{i,j}EBO}{c_i^{\text{h}}}$ for all $i \in I, j \in J$;
$(k, l) := \arg\max_{(i,j) \in I \times J} \Gamma_{i,j}$;
$\mathbf{S} := \mathbf{S} + e_{k,l}$.

Step 3. $C(\mathbf{S}) := C(\mathbf{S}) + c_k^{\text{h}}$;
Calculate $EBO_j(\mathbf{S})$ for all $j \in J^{\text{loc}}$;
If $EBO_j(\mathbf{S}) \le EBO_j^{\text{obj}}$ for all $j \in J^{\text{loc}}$, then stop, else go to step 2.

4.5. Alternative service measures

We discussed, in Section 3.6, alternative service measures for the single-location model. Analogous to the changes that we have to make for single-location models, we can make changes to the METRIC model to incorporate those service measures. However, we discuss only one of those service measures, the expected waiting time in Section 4.5.1. We then discuss usage of a service level constraint on the total expected number of backorders over all local warehouses in Section 4.5.2.

4.5.1. Expected waiting time

As in the single-location problem (see Section 3.6.1), it is straightforward to adapt the service level constraint based on the expected number of backorders to one that is based on the expected waiting time. This service measure has already been used, for example, by Hopp et al. [62] and Caglar et al. [63]. The expected waiting time for getting a ready-for-use part of LRU $i \in I$ at local warehouse $j \in J^{\text{loc}}$ when the base stock level is $S_{i,0}$ at the central depot and $S_{i,j}$ at the local warehouse, $W_{i,j}(S_{i,0}, S_{i,j})$, can be determined by Little's formula [56]: $W_{i,j}(S_{i,0}, S_{i,j}) = EBO_{i,j}(S_{i,0}, S_{i,j})/m_{i,j}$. Taking all LRUs together, the aggregate expected waiting time $W_j(\mathbf{S}_0, \mathbf{S}_j)$ at local warehouse $j \in J^{\text{loc}}$ is:

$$\begin{aligned} &W_j(\mathbf{S}_0, \mathbf{S}_j) \\ &\quad = \sum_{i \in I} \mathbb{P}\{\text{an arbitrary demand at local warehouse } j \text{ is for LRU } i\} \\ &\qquad \times (\text{expected waiting time for LRU } i \text{ at local warehouse } j) \\ &\quad = \sum_{i \in I} \frac{m_{i,j}}{\sum_{k \in I} m_{k,j}} \times \frac{EBO_{i,j}(S_{i,0}, S_{i,j})}{m_{i,j}} = \sum_{i \in I} \frac{EBO_{i,j}(S_{i,0}, S_{i,j})}{\sum_{k \in I} m_{k,j}}. \end{aligned}$$

This means that we get a problem that is equivalent to the original problem.

4.5.2. Expected number of backorders over all local warehouses

Instead of using a service level target per local warehouse, we can also use an aggregate service level target over all local warehouses:

$$\sum_{j \in J^{\text{loc}}} EBO_j(\mathbf{S}_0, \mathbf{S}_j) \leq EBO^{\text{obj}},$$

where EBO^{obj} denotes the aggregate target. This is the service level target that Sherbrooke uses in his seminal paper [16] and in his book [18]. Using such restriction can be interesting for users/maintenance organizations (see Section 2.1); OEMs servicing the installed base of their customers (see Section 2.2) have to achieve a certain service level for each customer.

Optimization can be done using a greedy algorithm that is very similar to Algorithm 3.1, but finding efficient solutions is not guaranteed. To find efficient solutions, a more time consuming greedy-like algorithm can be used, see [18,19] or [64]. The basic idea of that algorithm is to enumerate the stock level at the central depot, and to use, for each of those stock levels, a greedy algorithm to stock spare parts at the local warehouses. This leads to a number of sets of solutions. Their union leads to one set of solutions and the solutions that are efficient solutions for this set, are also efficient solutions for our optimization problem.

5. Emergency and lateral shipments

In several real-life situations, downtime of technical systems is too expensive to just patiently wait for a delivery when a requested spare part cannot be delivered immediately from the nearest local warehouse. In such situations, there are typically procedures in place to deliver a spare part from another source, especially in OEM networks, see Section 2.2. The part can be delivered from a neighboring local warehouse via a lateral transshipment or from the central depot via an emergency procedure. Alternatively, an emergency repair of the defective part is executed. Such procedures are typically expensive, but it is very beneficial to apply them if downtime costs of technical systems are much higher. For the local warehouse where the demand was placed initially, it means that the demand is lost instead of backordered when it cannot be satisfied from stock. (Therefore, the case with emergency shipments is equivalent with the lost sales case; for a review on lost sales models, including models specifically aimed at spare parts, see [65]). This has some consequences for our models and their analysis. We discuss this for single-location models in Section 5.1 and for two-echelon systems in Section 5.2. For the lateral transshipments, we restrict ourselves to *reactive* transshipments, although *proactive* transshipments have also been considered in the literature. The latter are transshipments that are performed periodically, e.g., at the beginning of each week. We refer to [66] for an overview of the literature on both reactive and proactive lateral transshipments. A recent paper in which the two types of transshipments are compared, and that contains an extensive overview of the related literature, is [67].

5.1. A single-location model with emergency shipments

In this section, we discuss a multi-item, single-location model with emergency shipments. We consider the minimization of total costs, consisting of inventory holding costs and costs for emergency shipments, subject to an aggregate mean waiting time constraint. This is the type of minimization problem that is relevant for the real-life networks as presented in Section 2. As mentioned above, the application of emergency shipments when demands arrive while the on-hand stock is zero, is equivalent to having lost sales. Already in 1957, Karush considered a similar, multi-item, single-location model with lost sales. In his model, he looked at the minimization of the total lost sales costs (lost revenues) under a given budget for the total stock. Also Feeney and Sherbrooke [49] already studied the lost sales case for their single-item model (see [68] for some corrections on their lost sales results).

Consider the single-location model of Section 3, but assume now that a demand is fulfilled from elsewhere by an emergency shipment in case a requested part cannot be fulfilled from stock. Assume that the average time for an emergency shipment is equal to t_i^{em} for LRU i. Instead of a constraint on the aggregate expected number of backorders, we get a constraint on the aggregate mean waiting time. Define $W_i(S_i)$ as the mean waiting time for an arbitrary demand for LRU i. It holds that:

$$W_i(S_i) = (1 - \beta_i(S_i))t_i^{\text{em}},$$

where $\beta_i(S_i)$ is the fill rate for LRU i. Next, we define $W(\mathbf{S})$ as the mean waiting time for an arbitrary demand for all LRUs together:

$$W(\mathbf{S}) = \sum_{i \in I} \frac{m_i}{M} W_i(S_i).$$

There are two types of costs now. As before, for each LRU i, there are unit inventory holding costs c_i^{h}. In addition, there are costs for the emergency shipments: c_i^{em}, each time that an emergency shipment is performed for LRU i. We assume that c_i^{em} contains the costs for a fast delivery from another location. When an emergency delivery is applied, one regular replenishment is applied less, and therefore those costs are subtracted. The average costs per time unit for LRU i for emergency shipments are equal to $m_i(1 - \beta_i(S_i))c_i^{\text{em}}$. This leads to the following formula for the total average costs per time unit for LRU i:

$$C_i(S_i) = c_i^{\text{h}} S_i + m_i(1 - \beta_i(S_i))c_i^{\text{em}},$$

and the total average costs, over all LRUs together, are equal to $C(\mathbf{S}) = \sum_{i \in I} C_i(S_i)$. The optimization problem, Problem (P′), that we want to solve is as follows:

$$\begin{aligned} \min \quad & C(\mathbf{S}) \\ \text{subject to} \quad & W(\mathbf{S}) \leq W^{\text{obj}} \\ & \mathbf{S} \in \mathcal{S}. \end{aligned} \tag{P′}$$

The evaluation of a given basestock policy $\mathbf{S}$ can still be done per LRU. Under the application of emergency shipments, the number

of parts in the repair pipeline of LRU i is limited from above by S_i. I.e., the behavior of the number of parts in repair of LRU i is no longer as in an $M/G/\infty$ queue but as in an $M/G/c/c$ queue, with $c = S_i$ parallel servers, arrival rate m_i, and mean service time t_i. The $M/G/c/c$ queue is also called an *Erlang loss system*. The fill rate $\beta_i(S_i)$ of LRU i is obtained via the Erlang loss probability. The fill rate is equal to the fraction of time that there is at least one part on stock, which is equal to the fraction of time that at least one server is free in the corresponding Erlang loss system. The latter probability is equal to 1 minus the fraction of time that all servers are occupied, i.e., to 1 minus the Erlang loss probability. Hence:

$$\beta_i(S_i) = 1 - \frac{(m_i t_i)^{S_i} \frac{1}{S_i!}}{\sum_{j=0}^{S_i} (m_i t_i)^j \frac{1}{j!}}. \tag{15}$$

Karush [69] has shown that the Erlang loss probability is strictly convex and decreasing as a function of the number of servers (see also Remark 2 in [70]). These properties imply that for each $i \in I$, $\beta_i(S_i)$ is strictly concave and increasing on its whole domain. As a result:

- For each $i \in I$, $W_i(S_i)$ is strictly convex and decreasing on its whole domain.
- For each $i \in I$, $C_i(S_i)$ is strictly convex on its whole domain. The function $C_i(S_i)$ is increasing for larger values of S_i and can be decreasing for smaller values of S_i because of the presence of the emergency costs.

Let $S_{i,\min} := \arg\min C_i(S_i)$. Then, obviously, for Problem (P′) and its corresponding multi-objective programming problem, we may exclude solutions with $S_i < S_{i,\min}$ for some $i \in I$. Then, efficient solutions can be generated for $C(\mathbf{S})$ and $W(\mathbf{S})$ in a similar way as for Problem (Q) in Section 3 (see [52, Section 8]); the factors Γ_i are now computed as $\Gamma_i := \Delta W_i(S_i)/\Delta C_i(S_i)$ with $\Delta W_i(S_i) = W_i(S_i) - W_i(S_i + 1)$ and $\Delta C_i(S_i) = C_i(S_i + 1) - C_i(S_i)$, $S_i \geq S_{i,\min}$.

5.2. Two-echelon models with lateral and emergency shipments

In this section, we discuss a two-echelon model with lateral and emergency shipments. Research on systems with lateral transshipments (proactive and reactive) started already in the late 1950s, with a first paper by Allen [71]; see [66]. The research on two-echelon systems with emergency shipments from the central warehouse (and a central repair facility) started much later with the work of Muckstadt and Thomas [72]. The first works with both emergency and lateral shipments are by Dada [73] and Alfredsson and Verrijdt [74]. The model described below can be seen as a multi-item version of the model of Alfredsson and Verrijdt [74]. For the optimization, the authors looked at a single-item minimization problem for the total costs consisting of inventory holding costs, costs for lateral and emergency shipments and costs for the delays when a demand is fulfilled via a lateral or emergency shipment. The model that we formulate is a multi-item model with explicit aggregate waiting time constraints, cf. [61,75]. For the discussion of solution methods, we focus on methods that are fast and accurate for systems with many items and local warehouses, or that form a basis for such methods.

Consider the two-echelon model of Section 4. We now assume that the following alternative options are considered to satisfy a demand for an LRU $i \in I$ at local warehouse $j \in J^{\text{loc}}$ if local warehouse j is out of stock:

1. *Lateral transshipment:* First, the stocks are checked at one or more other local warehouses $k \in J^{\text{loc}}$, $k \neq j$, that are at a relatively close distance to j. If one of these warehouses has a part on stock, then the demand is immediately coupled to that part and the part is delivered at the required place. This leads to a delay $t^{\text{lat}}_{i,j,k}$ and an extra cost $c^{\text{lat}}_{i,j,k}$ for the lateral transshipment itself.
2. *Emergency shipment from the central depot:* If the demand cannot be fulfilled by the above option, then the stock at the central depot is checked. If the central depot has a part on stock, then the demand is immediately coupled to that part and the part is delivered at the required place via an emergency shipment. This leads to a delay $t^{\text{cd}}_{i,j}$ and an extra cost $c^{\text{cd}}_{i,j}$ for the emergency shipment itself.
3. *Emergency shipment from the repair shop:* If the demand cannot be fulfilled by the above two options, then a ready-for-use part will be delivered from the repair shop. It is assumed that this is always possible, e.g., by quickly finishing the repair of a part of LRU i in the shop. This leads to a delay $t^{\text{rs}}_{i,j}$ and an extra cost $c^{\text{rs}}_{i,j}$ for the emergency repair/shipment itself.

Generally, the delays and extra costs increase when moving down in this list of options.

In this case, the use of the above options may depend on the basestock levels at all locations. Therefore, we define $\hat{\mathbf{S}}_i := (S_{i,0}, S_{i,1}, \ldots, S_{i,|J^{\text{loc}}|})$, $i \in I$; and, $\hat{\mathbf{S}} := (\hat{\mathbf{S}}_1, \ldots, \hat{\mathbf{S}}_{|I|})$. Let $\beta_{i,j}(\hat{\mathbf{S}}_i)$ be the probability that a demand for LRU i at local warehouse j is satisfied by the local warehouse itself. And let $\alpha_{i,j,k}(\hat{\mathbf{S}}_i)$, $\theta_{i,j}(\hat{\mathbf{S}}_i)$, and $\gamma_{i,j}(\hat{\mathbf{S}}_i)$ be the probabilities that the demand is satisfied by a lateral transshipment from another local warehouse k, an emergency shipment from the central depot, and an emergency shipment from the repair shop, respectively. It holds that:

$$\beta_{i,j}(\hat{\mathbf{S}}_i) + \sum_{k \in J^{\text{loc}}, k \neq j} \alpha_{i,j,k}(\hat{\mathbf{S}}_i) + \theta_{i,j}(\hat{\mathbf{S}}_i) + \gamma_{i,j}(\hat{\mathbf{S}}_i) = 1,$$
$$\forall i \in I,\ \forall j \in J^{\text{loc}}.$$

The mean waiting time for a demand of LRU i at local warehouse j is given by:

$$W_{i,j}(\hat{\mathbf{S}}_i) = \sum_{k \in J^{\text{loc}}, k \neq j} \alpha_{i,j,k}(\hat{\mathbf{S}}_i) t^{\text{lat}}_{i,j,k} + \theta_{i,j}(\hat{\mathbf{S}}_i) t^{\text{cd}}_{i,j} + \gamma_{i,j}(\hat{\mathbf{S}}_i) t^{\text{rs}}_{i,j},$$

and the aggregate mean waiting time for an arbitrary demand at local warehouse j is given by:

$$W_j(\hat{\mathbf{S}}) = \sum_{i \in I} \frac{m_{i,j}}{\sum_{k \in I} m_{k,j}} W_{i,j}(\hat{\mathbf{S}}_i).$$

For the total costs of LRU i, it holds that:

$$C_i(\hat{\mathbf{S}}_i) = c^{\text{h}}_i \sum_{j \in J} S_{i,j} + \sum_{j \in J^{\text{loc}}} m_{i,j} \times \left\{ \sum_{k \in J^{\text{loc}}, k \neq j} \alpha_{i,j,k}(\hat{\mathbf{S}}_i) c^{\text{lat}}_{i,j,k} + \theta_{i,j}(\hat{\mathbf{S}}_i) c^{\text{cd}}_{i,j} + \gamma_{i,j}(\hat{\mathbf{S}}_i) c^{\text{rs}}_{i,j} \right\}.$$

The total costs for all LRUs together are given by $C(\hat{\mathbf{S}}) = \sum_{i \in I} C_i(\hat{\mathbf{S}}_i)$.

Let W^{obj}_j denote the target for the aggregate mean waiting time at local warehouse j. Then, our optimization problem, Problem (R′), can be formulated as:

$$\begin{aligned} \min \quad & C(\hat{\mathbf{S}}) \\ \text{subject to} \quad & W_j(\hat{\mathbf{S}}) \leq W^{\text{obj}}_j, \quad \forall j \in J^{\text{loc}} \\ & S_{i,j} \in \mathbb{N}_0, \quad \forall i \in I,\ \forall j \in J. \end{aligned} \tag{R′}$$

For the Problems (P), (R), and (P′), as described in the Sections 3, 4, and 5.1, it is possible to do exact and efficient evaluations of given solutions. Unfortunately, this does not hold for Problem (R′). Hence, many approximate evaluation procedures have been developed for the above and related models. Based on an approximate evaluation procedure, good feasible solutions can be generated by greedy heuristics. Both the approximate evaluation procedures and the greedy heuristics are discussed below. We start the discussion for a system that is one step simpler.

Consider the above system, but with infinite stock in the central depot. We then get a *single-echelon, multi-location system*. An emergency shipment from the central depot will always be possible, and thus $\gamma_{i,j}(\hat{\mathbf{S}}_i) = 0$ for all $i \in I$ and $j \in J^{\text{loc}}$. The above expressions for $W_{i,j}(\hat{\mathbf{S}}_i)$, $W_j(\hat{\mathbf{S}})$, and $C(\hat{\mathbf{S}})$ are still valid (although the $S_{i,0}$ have to be excluded in the vectors $\hat{\mathbf{S}}_i$ and $\hat{\mathbf{S}}$). In the expression for $C_i(\hat{\mathbf{S}}_i)$, the inventory holding costs $c_i^{\text{h}} S_{i,0}$ at the central depot have to be excluded. The remaining problem is denoted as Problem (R'').

Problem (R'') has been studied by Kranenburg and Van Houtum [34]. The authors develop an iterative, approximate evaluation procedure based on a decomposition of the connected local warehouses into individual local warehouses and a modeling of the overflow demand streams (i.e., the requests for lateral transshipments) as Poisson demand streams with appropriate rates. Under this decomposition, the individual local warehouses have a steady-state behavior that is identical to that of Erlang loss systems (i.e., $M/G/c/c$ queues) with appropriately chosen parameters. As a result, simple closed-form expressions are obtained for this behavior. In the iterative algorithm, the steady-state behavior of the individual local warehouses and the rates of the overflow demand streams are alternatingly updated. This leads to an approximate evaluation procedure that has been shown to be accurate and efficient. The main idea behind this procedure stems from Axsäter [76]. The procedure was further refined by Van Wijk et al. [77], who modeled the overflow demand streams as interrupted Poisson processes (and who allowed hold back levels). More restricted models have also been proposed, for example by allowing lateral transshipments in one direction only (see, e.g., [78]). Reijnen et al. [79] model a general structure for the lateral transshipments that generalizes many other structures that have been studied in the literature.

For Problem (R''), Wong et al. [61] formulated a greedy heuristic, and tested this heuristic for instances with up to 50 LRUs and 4 local warehouses. They used an exact evaluation method based on Markov processes (cf. [75]). The greedy procedure starts with the solution $\hat{\mathbf{S}}_i = (0, \ldots, 0)$ and first, in a greedy fashion, parts are added until no further decrease of $C_i(\hat{\mathbf{S}}_i)$ is obtained (similarly as for Problem (P′), the costs $C_i(\hat{\mathbf{S}}_i)$ can be decreasing for small base stock levels because of the lateral and emergency shipment costs). Next, all combinations of LRUs and local warehouses are taken into account at the same time, and parts are added in a greedy fashion until a feasible solution is obtained (this step is similar to the greedy steps for Problem (R)). Next, a local search procedure was applied to further improve the current solution within the space of feasible solutions. Wong et al. [61] tested the generated heuristic solutions against a lower bound obtained by Lagrange relaxation and found small optimality gaps for problems with sufficiently many LRUs (at least 50 LRUs, say). Kranenburg and Van Houtum [34] followed the same greedy heuristic, but without a local search being added, and used the above approximate procedure for evaluation of given solutions. They compared the obtained solutions against a lower bound obtained by Dantzig–Wolfe decomposition and found also small optimality gaps.

Let us now return to Problem (R′). Intuitively, a greedy heuristic that is similar as for Problem (R'') should lead to good feasible solutions. In order to obtain a sufficiently fast procedure for problems of real-life size, such a greedy heuristic should be based on an efficient and still accurate approximate evaluation method. However, so far, such a method is not available. A method that is accurate, but not sufficiently fast is the one developed by Alfredsson and Verrijdt [74] for a system where all local warehouses look for lateral transshipments from all other local warehouse (full pooling). The authors first aggregate the stock of all local warehouses and analyze a two-dimensional Markov process to obtain the fraction of demands satisfied by emergency shipments from the central depot and repair shop. Next, the decomposition idea of Axsäter [76] is applied to analyze the local warehouses and to obtain the fractions of demand satisfied by the own local warehouse and by lateral transshipments. The method requires a numerical solution of the two-dimensional Markov processes, which requires relatively much computational effort.

It is not clear what the best method is to solve Problem (R′). The most straightforward way would be via developing an efficient and accurate approximate evaluation method for the full two-echelon system. This may be based on combining ideas from the above papers and similar approaches for related systems. In particular, use of a two-dimensional Markov process, as in [74], is to be avoided. Maybe, a coupling between the local warehouses and the central depot can be made as in approaches for two-echelon models without lateral, but with emergency shipments, see [72,80]. An alternative way is to decompose the two-echelon system in a single-location model for the central depot (cf. the model in Section 3) and a model for the local warehouses (cf. the single-echelon, multi-location model as discussed above), where an appropriate coupling is needed between these two models. For example, first, basestock levels for the central depot are determined, where then targets are assumed for the required service towards the local warehouses, and next, the basestock levels for the local warehouses are generated such that the total system performance meets the aggregate mean waiting time constraints. The latter approach has the advantage of being able to incorporate other aspects that play a role at the central depot level.

Finally, we would like to mention that similar approaches have been used for a few variants of the systems discussed above: (i) a two-echelon system as in Problem (R′) but without emergency shipments from the central depot and repair shop, see [76]; (ii) a two-echelon system as in Problem (R′) but with first looking for an emergency shipment from the central depot and then for a lateral transshipment, see [81]; (iii) a two-echelon system as in Problem (R′) but without lateral transshipments shipments, see [72,80]; (iv) a single-echelon, multi-location system as in Problem (R'') but without emergency shipments, see [23]. All these papers deal with single-item models and the authors have developed fast and accurate approximate evaluation procedures. The latter procedures are appropriate for being used in heuristic procedures for multi-item variants of these models with system-oriented service constraints.

6. Extensions

In this section, we discuss various extensions that can be made to spare parts models. In practice, many networks consist of more than two echelon levels. Furthermore, not only the components that are used during maintenance (LRUs) need to be stocked, but also their subcomponents (called SRUs). We discuss both extensions in Section 6.1. In Section 6.2, we discuss including condemnation, i.e., the property that a certain percentage of repairable components turns out not to be repairable, for instance, because a component has been repaired already a number of times. We discuss replacing the one-for-one replenishment policy by batching of replenishments in Section 6.3, and in Section 6.4, we discuss servicing various types of customers, i.e., multiple demand classes. Next, differences in criticality of components are discussed in Section 6.5: not all components cause immediate downtime of the technical system and they should thus be treated differently. Dynamic allocation rules that take into account the actual status of the inventory network when deciding on how to fulfill a demand are discussed in Section 6.6, and we discuss the usage of data that results from condition monitoring of technical systems in Section 6.7. Finally, we discuss extensions in which spare parts

inventory models are combined with models for other problems. We consider three such examples. We discuss the interaction with capacity planning in the repair shop in Section 6.8. In Section 6.9, we discuss the relation with locating stock points when designing an inventory network, and we discuss the relation with the Level Of Repair Analysis (LORA) problem in Section 6.10.

6.1. Multi-echelon systems and multi-indenture product structures

As denoted in Section 2, in practice, many networks consist of more than two echelon levels. Fortunately, as long as the network has a distribution structure (i.e., a tree structure), both the exact and approximate evaluation procedure for the METRIC method as described in Section 4.3 can easily be extended to multi-echelon networks, see [18,8], and the references therein. The same holds for the greedy heuristic, as discussed in Section 4.4. Including lateral and emergency shipments leads to the same (or even more) complications as we discussed for two-echelon models in Section 5.2.

Especially in user networks, as discussed in Section 2.1, it is typically necessary to not only stock Line Replaceable Units (LRUs) that are used during maintenance, but also their subcomponents, Stock Replaceable Units (SRUs). These SRUs are used to repair LRUs in a repair shop; we say that such SRUs are at the second indenture level in the product structure. If an SRU is not in stock when needed, the repair leadtime of the LRU is increased. The effect of these delays can be incorporated when setting planned repair leadtimes; it is then the responsibility of the repair shop to manage the stock of SRUs. Alternatively, an integrated planning model for the LRUs and SRUs can be used. It depends on the exact environment which option makes most sense, see [30].

Multi-indenture models can be used for the integrated planning of both LRUs and SRUs, and possibly of components at indenture levels below the SRUs. The first extension of the METRIC model to multi-indenture models is the MOD-METRIC model of Muckstadt [82]. Evaluating a stocking policy in a multi-indenture, multi-echelon model can be done similar to what we showed in Section 4.3, see, e.g., [8]. The greedy heuristic, as described in Section 4.4, can also still be used to find a good solution. To the best of our knowledge, there has been no research on lateral and emergency shipments in multi-indenture models.

In multi-echelon, multi-indenture models, it is also possible to incorporate that certain fractions of repairs are performed at one echelon level, whereas the rest is repaired at another echelon level (repairs can even be distributed over more than two echelon levels). However, combining this with lateral and emergency shipments makes the analysis of the resulting model probably very difficult. Models that include outsourcing a certain fraction of repairs or replacing a certain fraction of failed components by new components, are easier to analyze, and may be combined with lateral and emergency shipments. Such models are said to include condemnation, see Section 6.2.

6.2. Condemnation

We have thus far assumed that either all parts of a given LRU can be repaired (repairable parts) or that they all are discarded and replaced by a newly purchased component (consumable parts). In the case of a repairable part, we have assumed that such repairs are always successful. In practice, however, components fail due to various reasons. Some of the resulting defects are repairable, while other are not. Besides, many parts can only be repaired for a limited number of times, because their performance slowly decreases after each repair. Such phenomena are referred to as condemnation.

From a modeling point of view, condemnation can easily be incorporated, which was already noticed by Feeney and Sherbrooke [49]. The idea is to introduce a parameter r_i that represents the probability for a failed part of LRU i that it can be repaired. At the most upstream location, we distinguish a mean repair leadtime t_i^{rep} and a mean procurement leadtime t_i^{proc}. Then an arbitrary failed part leads to the arrival of a ready-for-use/new part at the most upstream location after an average leadtime $t_i = r_i t_i^{\text{rep}} + (1 - r_i) t_i^{\text{proc}}$. These t_i's are the leadtimes that can be used in the models of Sections 3–5, without changing anything else.

6.3. Batching

In the models of Sections 3–5, we assume one-for-one replenishments in the whole system. In local warehouses, this is generally justified because they receive consolidated replenishments from central depots. However, at the most upstream locations, failed parts are sent into repair or orders are placed at outside suppliers: then some form of batching may be desired. This is especially true in user networks, as discussed in Section 2.1. Reasons for using batching can be fixed setup costs for certain repair activities, fixed ordering and delivery costs that are charged by external suppliers, or pack sizes that are prescribed by suppliers. Applying the logic of the Economic Order Quantity rule denotes that generally one-for-one replenishments will make sense for the more expensive LRUs, which have high inventory holding costs and/or low demand rates. For less expensive components, however, it may be appropriate to use a fixed batch size Q, and thus to follow an (s, Q)-policy instead of a basestock policy (see also Section 2.1). For the single-location model of Section 3, we can handle an (s, Q)-policy if a deterministic replenishment leadtime is assumed. For an LRU i with reorder level s_i and batch size Q_i, it then holds that at an arbitrary time point t, the inventory position (the amount of stock on hand plus on order) is uniformly distributed on the integers $s_i + 1, \ldots, s_i + Q_i$ (see, e.g., Proposition 5.1 of Axsäter [83]). This property can be used to determine the probability distribution at time point $t + t_i$, and the rest of the analysis goes along the same lines as before. For the two-echelon model of Section 4, the extension can be made in a similar way, see also [84]. For the models of Section 5 with emergency and lateral shipments, it is much less clear how the effect of a fixed batch size can be included.

6.4. Multiple demand classes

In many real-life OEM networks (see Section 2.2), some customers demand a higher service level than others, e.g., delivery after four hours versus next day delivery. There could exist, for instance, silver, gold, and platinum contracts that are increasingly higher priced.

The most simple way to deal with this, is to not differentiate at all in the spare parts supply and give all customers the highest service level. Although this policy, referred to as the 'round-up policy' (see, e.g., [85]), seems strange, it is regularly used in practice (see, e.g., [21]). One reason for this can be that management decides that the simplicity of this policy makes it attractive (it is still possible to differentiate customers in another way, e.g., by adapting the dispatching policy for service engineers). Another reason can be that service engineers see that there is stock on hand and do not want to let their customer wait, even though the customer has a low priced contract. The most simple way to differentiate between customer classes is by holding separate stocks for both customer classes. This policy is denoted as the 'separate stock' policy (see, e.g., [85]). The downside of this policy is that inventory pooling benefits are not capitalized, but, again, the policy can be interesting because of its simplicity to implement.

By far the most studied policy that combines demand streams of various customer classes while differentiating between the various classes, is the critical level policy (cf. [86]). The basic idea of this

policy is, for the example of two demand classes, that there is a certain critical level K_i for LRU i. When on hand stock is higher than or equal to K_i, demands from both classes of customers are fulfilled FCFS, but when on hand stock is lower than K_i, demands of premium customers are fulfilled only, while the demands of the other customers are backordered (or lost, depending on the model assumptions). This policy has been applied to a spare parts inventory setting in several papers, see, e.g., [85,87–90]. In these studies, the exact cost performance of a critical level policy is obtained via Markov/queueing analysis. Generally, the goal is to find the best policy within the class of critical level policies (via exact and heuristic methods). The cost of the best critical level policy has also been compared numerically to the round-up and separate stock policies, see, e.g., [85,88].

To the best of our knowledge, Abouee-Mehrizi et al. [91] are the only ones who consider multiple demand classes in a two-echelon network, i.e., a network consisting of local warehouses with, possibly, different backordering (and holding) costs, and a central warehouse with finite inventory. Using at the central warehouse a critical level policy as described above, could lead to sending a spare part to a high priority local warehouse (having high backordering costs) that already holds on-hand stock, while a low priority local warehouse is out of stock. Abouee-Mehrizi et al. compare various priority rules at the central depot, including the critical level policy and a generalized version of that policy (which they call the Multilevel Rationing and Generalized Multilevel Rationing policies, respectively).

There are also other ways to differentiate between customer classes. Alvarez et al. [92] for instance, propose to use, in the case of a stock out, emergency shipments for priority customers only, and Alvarez et al. [93] propose to differentiate by using dedicated stocks at some customers only.

6.5. Criticality

So far, we have assumed that if any component fails, the entire system fails. In practice, this assumption is not always justified; there are many components whose failures do not lead to the system breaking down. We say that the criticality of a component need not be 100%: the criticality of a part is related to the consequences for the system if that part is not replaced immediately. When an OEM has a service contract with the users of its systems, then this criticality may not be relevant. However, for users that maintain their own systems, criticality is typically very relevant to consider, as discussed in Section 2.1.

The criticality can depend on the part itself (e.g., a toilet in a train is not critical), the exact failure mode of a failed part, the position of a component in the system (a component can occur at multiple places in a system, and the criticality may differ per position), the level of redundancy per position, and so on. A good way to address these factors is going back to the reliability data of a system and incorporate the above factors in the modeling. This has been done, for instance, by Van Jaarsveld and Dekker [94]. De Smidt-Destombes et al. have proposed spare parts models for k-out-of-n systems in which there are n identical components, out of which k ($k < n$) need to function.

6.6. Dynamic allocation rules

In the models discussed in Section 5, the rules for the use of emergency and lateral shipments are denoted as static rules. This is appropriate for inventory models that are used at a tactical level. At the operational planning level, however, it can be more beneficial to use dynamic rules, under which the choice of the option to fulfill a demand depends on the actual status of the inventory network. Suppose that a demand occurs at a local warehouse that is out of stock at that moment. Then it may make sense to look at when a next part arrives from the replenishment pipeline. If a part would arrive in half a day, and a lateral transshipment would take an equal or longer time, then it makes sense to fulfill the demand by the part in the pipeline. Such usage of pipeline information has been shown to be quite beneficial in certain cases, e.g., by Chu et al. [95], Axsäter [96], Minner et al. [97], Howard et al. [44] and Yang et al. [98].

Chu et al. [95] focus on a single stock point and consider a policy of backordering demands that cannot be fulfilled immediately up to a certain backorder limit and losing sales after that. After a certain time point, the backorder limit can be increased, thus utilizing pipeline information. The authors use results from regenerative processes and from [99] to derive the cost function. An exhaustive search is required to find the optimal solution.

The other four papers consider a single-echelon, multi-location model (see Section 5.2). A demand that cannot be covered (completely) from on-hand stock can be covered (partly) by a lateral transshipment, except in the model of Howard et al. [44], where an emergency shipment can be used (see below). Any remaining demand is backordered. The decision on whether or not to use a lateral transshipment (or emergency shipment) is made incorporating pipeline information. Axsäter [96] determines the decision rule myopically by choosing the best option from a set of alternatives, assuming that no further transshipments will take place. The difference in cost resulting from choosing each option is compared with the long-run average costs. Minner et al. [97] compare, for each possible transshipment size, the approximated resulting costs and then perform a search over those possible sizes. Howard et al. [44] consider a slightly different network, to which also a support warehouse is added. The authors use the overflow modeling introduced by Axsäter [76, and explained for Problem R'' in Section 5.2] to decompose the problem into smaller problems. Queueing theory is applied in order to determine a good policy for ordering emergency shipments. Yang et al. [98] analyze a queueing model that approximates the behavior of the system that they have modeled and they use the above mentioned procedure of Axsäter [76].

Another type of information that can be beneficial to use, is about the actual on hand stocks of neighboring local warehouses when a lateral transshipment is needed. It then makes sense to fulfill a demand from a neighboring local warehouse that has still two or more parts on stock rather than from a warehouse that has only one part left, see, e.g., [100,101].

6.7. Advance demand information

Inventory control may benefit from using information on the condition of the installed base and from other forms of advance demand information (ADI), e.g., resulting from customers that place orders that will lead to an actual demand only after a certain lead time. The seminal paper on ADI is that of Hariharan and Zipkin [102]. The authors assume a continuous review, base stock policy with full backordering in both a single location system and a serial system. Through construction of equivalent systems and using known results from the literature, for example on Poisson processes, Hariharan and Zipkin [102, p. 1600] find that "the effect of a demand leadtime on overall system performance is precisely the same as a corresponding reduction in the supply system".

We are aware of the following papers that consider, in the context of spare parts inventory control, ADI that results from monitoring the degradation state of components in the installed base, all of which consider a periodic review model: [3,103–105]. The papers differ greatly in modeling assumptions and thus in the resulting analysis. Key are the differences in the demand process. Deshpande et al. [3] assume that a part-age signal can be observed, which is compared with a threshold. Depending on the number of

parts that have a signal above the threshold value, the authors calculate a conditional mean and variance of a normally distributed lead time demand. Li and Ryan [103] model deterioration of each part as a Wiener process and use that to estimate the distribution of the remaining useful life of each part. The estimate is updated each period using Bayesian updating and it is used to estimate the distribution of the demand for spare parts in the upcoming periods. Louit et al. [104] use the proportional hazards model (PHM) to model the hazard rate function of one item in one machine. This means that they assume a deterministic hazard function that depends on the age only, plus a vector of time-dependent covariates that result from condition monitoring. They use it to get an estimate of the conditional reliability and the remaining useful life of the item. Lin et al. [105] assume that there are multiple machines, each with one critical item. Each of the items' degradation processes are modeled as a Markov chain. This leads to a multi-dimensional Markov decision process.

In addition to adapting the spare parts supply, the maintenance policy can also be adapted, see, e.g., [106–109].

6.8. Repair shop capacity planning

Decisions that are made about the repair shop, heavily influence the amount of spare parts that is required; enabling faster repairs can mean that less spare parts are required to achieve the same service level. In some cases, repairs themselves become faster, for example, by automating part of the process; in other cases, delays in the repair shop due to queueing in front of bottlenecks is reduced by installing more equipment or hiring more manpower (more servers). Especially in the case of a user who maintains its own systems (see Section 2.1), decisions on the repair shop and the inventory control can be made jointly, which can lead to lower costs than when the inventory control policy is determined after the decisions on the repair shop have been made.

To the best of our knowledge, Van der Heijden et al. [110] have been the only ones to consider the explicit trade-off between enabling faster repairs and stocking more spare parts on a tactical level. For a multi-echelon, multi-indenture model, they extend the greedy heuristic that we have discussed in Sections 3.4.2 and 4.4 to include the option to reduce certain lead times, instead of stocking more spare parts. Various authors have considered the trade-off between adding more servers (e.g., equipment) and adding more spare parts: there exist various papers in which capacitated repair shops have been introduced in METRIC type models (see, e.g., [111]) and there has also been some work on the joint optimization of the number of servers and amount of spare parts (see, e.g., [112]).

6.9. Facility location problem

Especially for OEMs that perform maintenance on the installed base that they have sold (see Section 2.2), it can be worthwhile to not see the layout of the spare parts network as a given. Companies such as IBM increasingly depend on logistics service providers (LSPs) such as DHL or UPS to stock their spare parts close to the customers. In other words, LSPs are responsible for the stocking locations at echelon level 1 (see Fig. 2). As a result, it is relatively inexpensive to change the stocking locations and the decision on where to locate these stock points, the facility location problem, thus becomes a tactical problem instead of a strategic problem.

The locations of the stock points and the amount of spare parts may then be optimized jointly, which is especially relevant if time based service constraints have been agreed upon with customers (e.g., a certain percentage of requested parts should be fulfilled within a certain number of hours). Candas and Kutanoglu [113], for instance, solve the joint problem by linearizing fill rate functions and then solving the resulting mixed integer linear programming formulation. Rappold and Van Roo [114] use a two-step approach: the first step consists mainly of the facility location problem whereas the second step considers the spare parts inventories. They take the capacities of the locations into account in both steps, which Candas and Kutanoglu, for example, do not do.

6.10. Level of repair analysis

Level Of Repair Analysis (LORA) is used to decide on (1) which components to repair upon failure and which to discard, (2) where in the repair network to perform repairs, and (3) where to install the capabilities that are required to perform repairs (and possibly discards), such as manpower and equipment. In the models that we have discussed above, we have implicitly assumed that such decisions have already been taken. However, these decisions need to be made carefully in practice since they influence the required amount of spare parts drastically. For instance, repairing a certain component at echelon level 1 leads to a lead time that is short compared to discarding that component and replacing it by a newly purchased component. This lead time determines to a large extent the required amount of spare parts. LORA is a term that stems from the military. It is a key analysis in Logistics Support Analysis (LSA), which is a process used to increase efficiency of future maintenance throughout the development process of new (military) equipment. More on LORA and LSA can be found, for example, in [115]. Because of the importance of the LORA problem, a number of papers has been written on this subject.

Barros [116] and Saranga and Dinesh Kumar [117] propose mixed integer linear programming (MILP) formulations to model the LORA problem. Basten et al. [118] propose a MILP formulation that generalizes those two formulations and Basten et al. [119] propose an improved formulation that has similarities with a minimum cost flow model and can therefore be solved efficiently (using a standard solver, e.g., CPLEX). Basten et al. [120] model various extensions to the LORA problem, such as the occurrence of unsuccessful repairs or no-fault-founds. Brick and Uchoa [121] propose a model that is similar to that of [119]. The key difference is that the former integrate in the LORA the problem of deciding which facilities to open out of a set of possible facilities. Furthermore, Brick and Uchoa consider one echelon level, effectively assume two indenture levels and assume that resources are capacitated.

Alfredsson [122] and Basten et al. [64] solve the joint problem of LORA and spare parts stocking integrally. However, in order to do so, they have to make simplifying assumptions. For example, they consider two-echelon networks only and single-indenture product structures. Basten et al. [123] focus on more general network and product structures, but they have to resort to a two-step approach. In the first step, they perform the LORA, and in the second step, they take the decisions on spare parts inventories. They next feedback the results on the spare parts inventories to the LORA and they generally find good solutions after a couple of iterations. The drawback is that there is no guarantee on the performance of their heuristic. Basten et al. [64,123] show that solving the integrated problem instead of first solving the LORA problem and then the spare parts stocking problem, leads to cost reductions of over 3% on average and over 30% at maximum.

7. Conclusions and application in practice

We have discussed a broad range of spare parts inventory control models. It should be clear by now that there have been a lot of results over, approximately, the last 50 years, and that there are still a lot of opportunities for further research on these models. Combining spare parts inventory control problems with other problems poses interesting challenges too, of which some have

been treated in Section 6. Some of the models that we have discussed have found their way into applications that are used in practice. We end this survey paper by pointing to some literature that presents these applications.

Multi-echelon, multi-indenture, METRIC-type models have been implemented in various software tools, in use at various organizations. Historically, military organizations, especially the US military forces, have been the first to adopt such models. In his book, Sherbrooke [18, Chapter 10] discusses extensively that various models (e.g., VARI-METRIC and MOD-METRIC) have been implemented in the US military forces decades ago. For instance, VMetric (see, e.g., [18, Appendix E]) is a software tool that is based on VARI-METRIC and that has been used by, among others, the US Coast Guard; VMetric has been developed by TFD Group.

OPUS10, by Systecon, has been used by various European aircraft manufacturers and NATO air forces [18, p. 232]; Inventri, based on VARI-METRIC and the work of Rustenburg [28], is in use at Thales Nederland [120].

Also outside the military, these models have been applied. Cohen et al. [37] report on a software tool in use at IBM, Optimizer, that uses multi-indenture, multi-echelon models. Morris Cohen is further one of the founders of MCA Solutions (Morris Cohen & Associates). Both MCA Solutions and Xelus, which offered Xelus Parts, have been acquired by Servigistics, which offers Servigistics' Service Parts Management. All these software tools are (were) based on multi-indenture, multi-echelon models.

Finally, some other models have found their way into applications as well. A single-echelon, multi-location model with lateral transshipments (mentioned at the end of Section 5.2) has been implemented at ASML, see, e.g., [34]. Deshpande et al. [3] report on a project performed at the US Coast Guard, in which, among other things, the usage of an inventory control policy that uses advance demand information (see Section 6.7) is investigated. At the time of writing by Deshpande et al., the US Coast Guard was contracting commercial vendors to develop decision support containing this policy.

Draper and Suanet [124] explain that IBM has developed the Global Part System (GPS), which consists of both components that IBM has developed itself and components that it has purchased. At the time of writing by Draper and Suanet, GPS was used to support IBM's four central warehouses only, but it was planned that its patented Network Neighborhood would be integrated in order to support local stock points. Network Neighborhood uses time based service level targets and lateral transshipments [125].

Acknowledgments

The authors thank two anonymous reviewers and the associate editor for their valuable feedback, which improved the original paper. The first author gratefully acknowledges the support of the Lloyd's Register Foundation (LRF). LRF helps to protect life and property by supporting engineering-related education, public engagement and the application of research.

References

[1] B. Kranenburg, Spare Parts Inventory Control Under System Availability Constraints (Ph.D. thesis), BETA Research School, Eindhoven, The Netherlands, D88, 2006.
[2] L. Harrington, From just in case to just in time, Air Transp. World (2007) 77–80.
[3] V. Deshpande, A.V. Iyer, R. Cho, Efficient supply chain management at the U.S. Coast Guard using part-age dependent supply replenishment policies, Oper. Res. 54 (6) (2006) 1028–1040.
[4] R.J.I. Basten, Designing Logistics Support Systems. Level of Repair Analysis and Spare Parts Inventories (Ph.D. thesis), BETA Research School, Enschede, The Netherlands, D128, 2010.
[5] M.A. Cohen, Y.-S. Zheng, V. Agrawal, Service parts logistics: a benchmark analysis, IIE Trans. 29 (1997) 627–639.
[6] U.W. Thonemann, A.O. Brown, W.H. Hausman, Easy quantification of improved spare parts inventory policies, Manag. Sci. 49 (9) (2002) 1213–1225.
[7] W. Rustenburg, G. van Houtum, W. Zijm, Spare parts management at complex technology-based organizations: an agenda for research, Int. J. Prod. Econ. 71 (2001) 177–193.
[8] W.D. Rustenburg, G.J. Van Houtum, W.H.M. Zijm, Exact and approximate analysis of multi-echelon, multi-indenture spare parts systems with commonality, in: J.G. Shanthikumar, D.D. Yao, W.H.M. Zijm (Eds.), Stochastic Modelling and Optimization of Manufacturing Systems and Supply Chains, Kluwer, Boston, MA, 2003, pp. 143–176.
[9] Thales, Extended services. A portfolio offering total solutions for all customer needs, 2011. Retrieved from http://www.thalesgroup.com/Portfolio/Documents/Extended_Services_(February_2011)/?LangType=2057, last checked on February 2, 2012.
[10] K.R. Montgomery, C.B. Thorstenson, Life Cycle Costs of Alternatives for F-16 Printed Circuit Board Diagnosis Equipment (Master's thesis), Air force institute of technology, Dayton, OH, 1994.
[11] R. Dekker, Ç. Pinçe, R. Zuidwijk, M.N. Jalil, On the use of installed base information for spare parts logistics: a review of ideas and industry practice, Int. J. Prod. Econ. 143 (2) (2013) 536–545.
[12] R.H. Teunter, L. Fortuin, End-of-life service, Int. J. Prod. Econ. 59 (1999) 487–497.
[13] M. Van der Heijden, B.P. Iskandar, Last time buy decisions for products sold under warranty, European J. Oper. Res. 224 (2) (2013) 302–312.
[14] J.E. Boylan, A.A. Syntetos, Spare parts management: a review of forecasting research and extensions, IMA J. Manag. Math. 21 (2010) 227–237.
[15] R.J.I. Basten, E. van Wingerden, R. Dekker, W.D. Rustenburg, More grip on inventory control through improved forecasting. A comparative study at three companies, ERIM Research Paper Series 2012-24, 2012.
[16] C.C. Sherbrooke, METRIC: a multi-echelon technique for recoverable item control, Oper. Res. 16 (1) (1968) 122–141.
[17] C.C. Sherbrooke, Optimal Inventory Modeling of Systems. Multi-Echelon Techniques, first ed., Wiley, 1992.
[18] C.C. Sherbrooke, Optimal Inventory Modelling of Systems. Multi-echelon Techniques, second ed., Kluwer, Dordrecht, The Netherlands, 2004.
[19] J.A. Muckstadt, Analysis and Algorithms for Service Parts Supply Chains, Springer, New York, NY, 2005.
[20] S.-H. Kim, M.A. Cohen, S. Netessine, Performance contracting in after-sales service supply chains, Manag. Sci. 53 (12) (2007) 1843–1858.
[21] M.A. Cohen, N. Agrawal, V. Agrawal, Winning in the aftermarket, Harv. Bus. Rev. 84 (5) (2006) 129–138.
[22] Atos Consulting, Servitization in product companies. Creating business value beyond products, White Paper, 2011.
[23] A. Kukreja, C.P. Schmidt, D.M. Miller, Stocking decisions for low-usage items in a multilocation inventory system, Manag. Sci. 47 (10) (2001) 1371–1383.
[24] A. Cesaro, D. Pacciarelli, Performance assessment for single echelon airport spare part management, Comput. Ind. Eng. 61 (2011) 150–160.
[25] J. Van Duren, Differentiated Spare Parts Management: An Application in the Aircraft Industry (Master's thesis), Eindhoven University of Technology, Eindhoven, The Netherlands, 2011.
[26] R.A.M. Kusters, The Design of a Logistic Support System (Master's thesis), Eindhoven University of Technology, Eindhoven, The Netherlands, 2011.
[27] M. Braglia, M. Frosolini, Virtual pooled inventories for equipment-intensive industries. an implementation in a paper district, Reliab. Eng. Syst. Saf. 112 (2013) 26–37.
[28] W. Rustenburg, A System Approach to Budget-constrained Spare Parts (Ph.D. thesis), BETA Research School, Eindhoven, The Netherlands, D36, 2000.
[29] W. Rustenburg, G. Van Houtum, W. Zijm, Spare parts management for technical systems: resupply of spare parts under limited budgets, IIE Trans. 32 (2000) 1013–1026.
[30] M.A. Driessen, J.J. Arts, G.J. van Houtum, W. Rustenburg, B. Huisman, Maintenance spare parts planning and control: a framework for control and agenda for future research, Prod. Plan. Control (2014) in press.
[31] L.A.M. Van Dongen, Maintenance Engineering: Maintaining Links. Inaugural Lecture, University of Twente, Enschede, The Netherlands, 2011.
[32] T. Tinga, Application of physical failure models to enable usage and load based maintenance, Reliab. Eng. Syst. Saf. 95 (2010) 1061–1075.
[33] K.S. De Smidt-Destombes, M.C. van der Heijden, A. van Harten, On the availability of a k-out-of-n system given limited spares and repair capacity under a condition based maintenance strategy, Reliab. Eng. Syst. Saf. 83 (1) (2004) 287–300.
[34] B. Kranenburg, G.J. van Houtum, A new partial pooling structure for spare parts networks, European J. Oper. Res. 199 (2009) 908–921.
[35] I.M.H. Vliegen, Integrated Planning for Service Tools and Spare Parts for Capital Goods (Ph.D. thesis), BETA Research School, Eindhoven, The Netherlands, D123, 2009.
[36] R.M.J. Arts, Design of Spare Parts Network at Cisco Systems (Master's thesis), Eindhoven University of Technology, Eindhoven, The Netherlands, 2010.
[37] M.A. Cohen, P.V. Kamesan, P. Kleindorfer, H. Lee, A. Tekerian, Optimizer: IBM's multi-echelon inventory system for managing serive logistics, Interfaces 20 (1) (1990) 65–82.
[38] E. Kutanoglu, Insights into inventory sharing in service parts logistics systems with time-based service levels, Comput. Ind. Eng. 54 (2008) 341–358.
[39] M.N. Jalil, R. Zuidwijk, M. Fleischmann, J.A. van Nunen, Spare parts logistics and installed base information, J. Oper. Res. Soc. 62 (3) (2010) 442–457.

[40] P. Rijk, Multi-item, Multi-location Stock Control with Capacity Constraints for the Fieldstock of Service Parts at Océ (Master's thesis), Eindhoven University of Technology, Eindhoven, The Netherlands, 2007.
[41] R.M.A. Schetters, Redesigning the Spare Parts Network at Océ: a Scenario Analysis for Optimizing Total Relevant Network Cost from External Suppliers to Regional Warehouses (Master's thesis), Eindhoven University of Technology, Eindhoven, The Netherlands, 2010.
[42] M.A. Cohen, Y.-S. Zheng, Y. Wang, Identifying opportunities for improving teradynes service-parts logistics system, Interfaces 29 (4) (1999) 1–18.
[43] R.P. Vlasblom, Steering Life Cycle Costs in the Early Design Phase (Master's thesis), Eindhoven, The Netherlands, 2009.
[44] C. Howard, I. Reijnen, J. Marklund, T. Tan, Using pipeline information in a multi-echelon spare parts inventory system, BETA Working Paper 330, 2010.
[45] R. Oliva, R. Kallenberg, Managing the transition from products to services, Int. J. Serv. Ind. Manage. 14 (2) (2003) 160–172.
[46] Aberdeen Group, The Service Parts Management Solution Selection Report. SPM Strategy and Technology Selection Handbook, in: Service Chain Management. Featured Research Series., Aberdeen Group, Boston, MA, 2005.
[47] A. van Wijk, I. Adan, G.J. van Houtum, Optimal allocation policy for a multi-location inventory system with a quick response warehouse, Oper. Res. Lett. 41 (3) (2013) 305–310.
[48] S. Axsäter, C. Howard, J. Marklund, A distribution inventory model with transshipments from a support warehouse, IIE Trans. 45 (3) (2013) 309–322.
[49] G.J. Feeney, C.C. Sherbrooke, The $(s-1, s)$ inventory policy under compound Poisson demand, Manag. Sci. 12 (5) (1966) 391–411.
[50] C. Palm, Analysis of the erlang traffic formulae for busy-signal arrangements, Ericsson Tech. 4 (1938) 39–58.
[51] G. Van Houtum, K. Hoen, Single-Location, Multi-Item Inventory Models for Spare Parts, in: LNMB, 2008, Handout for Course 'Inventory Management in Supply Chains'.
[52] B. Fox, Discrete optimization via marginal analysis, Manag. Sci. 13 (3) (1966) 210–216.
[53] E.L. Porteus, Foundations of Stochastic Inventory Theory, Stanford University Press, Palo Alto, CA, 2002.
[54] H.I. Everett, Generalized lagrange multiplier method for solving problems of optimum allocation of resources, Oper. Res. 11 (1963) 399–417.
[55] G.B. Dantzig, P. Wolfe, Decomposition principle for linear programs, Oper. Res. 8 (1960) 101–111.
[56] J.D.C. Little, A proof for the queuing formula: $L = \lambda W$, Oper. Res. 9 (3) (1961) 383–387.
[57] G.J. Feeney, J.W. Petersen, C.C. Sherbrooke, An Aggregate Base Stockage Policy for Recoverable Spare Parts, The Rand Corporation, Santa Monica, CA, Rand Memorandum 3644-PR, 1963.
[58] G. Gallego, Ö. Özer, P.H. Zipkin, Bounds, heuristics, and approximations for distribution systems, Oper. Res. 55 (3) (2007) 503–517.
[59] S.C. Graves, A multi-echelon inventory model for a repairable item with one-for-one replenishment, Manag. Sci. 31 (10) (1985) 1247–1256.
[60] H. Wong, B. Kranenburg, G.J. van Houtum, D. Cattrysse, Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse, OR Spectrum 29 (2007) 699–722.
[61] H. Wong, G.J. van Houtum, D. Cattrysse, D. van Oudheusden, Simple, efficient heuristics for multi-item multi-location spare parts systems with lateral transshipments and waiting time constraints, J. Oper. Res. Soc. 56 (2005) 1419–1430.
[62] W.J. Hopp, R.Q. Zhang, M.L. Spearman, An easily implementable hierarchical heuristic for a two-echelon spare parts distribution system, IIE Trans. 31 (1999) 977–988.
[63] D. Caglar, C. Li, D. Simchi-Levi, Two-echelon spare parts inventory system subject to a service constraint, IIE Trans. 36 (2004) 655–666.
[64] R.J.I. Basten, M.C. Van der Heijden, J.M.J. Schutten, Joint optimization of level of repair analysis and spare parts stocks, European J. Oper. Res. 222 (3) (2012) 474–483.
[65] M. Bijvank, I.F.A. Vis, Lost-sales inventory theory: a review, European J. Oper. Res. 215 (2011) 1–13.
[66] C. Paterson, G. Kiesmüller, R. Teunter, K. Glazebrook, Inventory models with lateral transshipments: a review, European J. Oper. Res. 210 (2) (2011) 125–136.
[67] A. Seidscher, S. Minner, A semi-markov decision problem for proactive and reactive transshipments between multiple warehouses, European J. Oper. Res. 230 (1) (2013) 42–52.
[68] J. Chen, P.L. Jackson, J.A. Muckstadt, Exact analysis of a lost sales model under stuttering Poisson, Oper. Res. 59 (1) (2011) 249–253.
[69] W. Karush, A queueing model for an inventory problem, Oper. Res. 5 (1957) 693–703.
[70] B. Kranenburg, G.J. van Houtum, Cost optimization in the $(S-1, S)$ lost sales inventory model with multiple demand classes, Oper. Res. Lett. 35 (2007) 493–502.
[71] S.G. Allen, Redistribution of total stock over several user locations, Nav. Res. Logist. Q. (4) (1958) 51–59.
[72] J.A. Muckstadt, L.J. Thomas, Are multi-echelon inventory methods worth implementing in systems with low-demand-rate items? Manag. Sci. 26 (5) (1980) 483–494.
[73] M. Dada, A two-echelon inventory system with priority shipments, Manag. Sci. 38 (8) (1992) 1140–1153.
[74] P. Alfredsson, J. Verrijdt, Modeling emergency supply flexibility in a two-echelon inventory system, Manag. Sci. 45 (10) (1999) 1416–1431.
[75] H. Wong, G. van Houtum, D. Cattrysse, D. van Oudheusden, Multi-item spare parts systems with lateral transshipments and waiting time constraints, European J. Oper. Res. 171 (2006) 1071–1093.
[76] S. Axsäter, Modelling emergency lateral transshipments in inventory systems, Manag. Sci. 36 (11) (1990) 1329–1338.
[77] A. van Wijk, I. Adan, G.J. van Houtum, Approximate evaluation of multi-location inventory models with lateral transshipments and hold back levels, European J. Oper. Res. 218 (3) (2012) 624–635.
[78] F. Olsson, An inventory model with unidirectional lateral transshipments, European J. Oper. Res. 200 (2010) 725–732.
[79] I. Reijnen, T. Tan, G.J. van Houtum, Inventory planning for spare parts networks with delivery time requirements, BETA Working Paper 280, 2009.
[80] E. Özkan, G.J. van Houtum, Y. Serin, A new approximate evaluation method for two-echelon inventory systems with emergency shipments, Ann. Oper. Res. (2013). http://dx.doi.org/10.1007/s10479-013-1401-9. in press.
[81] J. Grahovac, A. Chakravarty, Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items, Manag. Sci. 47 (4) (2001) 579–594.
[82] J.A. Muckstadt, A model for a multi-item, multi-echelon, multi-indenture inventory system, Manag. Sci. 20 (4) (1973) 472–481.
[83] S. Axsäter, Inventory Control, second ed., Springer, New York, NY, 2006.
[84] E. Topan, Z.P. Bayındır, T. Tan, An exact solution procedure for multi-item two-echelon spare parts inventory control problem with batch ordering in the central warehouse, Oper. Res. Lett. 38 (5) (2010) 454–461.
[85] V. Deshpande, M.A. Cohen, K. Donohue, A threshold inventory rationing policy for service-differentiated demand classes, Manag. Sci. 49 (6) (2003) 683–703.
[86] D.M. Topkis, Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes, Manag. Sci. 15 (3) (1968) 160–176.
[87] R. Dekker, R.M. Hill, M.J. Kleijn, R. Teunter, On the $(s-1, s)$ lost sales inventory model with priority demand classes, Nav. Res. Logist. 49 (6) (2002) 593–610.
[88] A.A. Kranenburg, G.J. van Houtum, Service differentiation in spare parts inventory management, J. Oper. Res. Soc. 59 (2008) 946–955.
[89] K.T. Möllering, U.W. Thonemann, An optimal constant level rationing policy under service level constraints, OR Spectrum 32 (2) (2010) 319–341.
[90] P. Enders, I. Adan, A. Scheller-Wolf, G.J. van Houtum, Inventory rationing for a system with heterogeneous customer classes, Flexible Serv. Manuf. J. (2013). http://dx.doi.org/10.1007/s10696-012-9148-1. in press.
[91] H. Abouee-Mehrizi, O. Baron, O. Berman, Customer differentiation in capacitated multi-echelon inventory systems, Working Paper, 2012.
[92] E. Alvarez, M.C. Van der Heijden, W. Zijm, The selective use of emergency shipments for service-contract differentiation, Int. J. Prod. Econ. 143 (2) (2013) 518–526. http://dx.doi.org/10.1016/j.ijpe.2012.02.019.
[93] E.M. Alvarez, M.C. Van der Heijden, W. Zijm, Service differentiation in spare parts supply through dedicated stocks, Ann. Oper. Res. (2013). http://dx.doi.org/10.1007/s10479-013-1362-z. in press.
[94] W. Van Jaarsveld, R. Dekker, Spare parts stock control for redundant systems using reliability centered maintenance data, Reliab. Eng. Syst. Saf. 96 (2011) 1576–1586.
[95] C.W. Chu, B.E. Patuwo, A. Mehrez, G. Rabinowitz, A dynamic two-segment partial backorder control of (r, q) inventory system, Comput. Oper. Res. 28 (2001) 935–953.
[96] S. Axsäter, A new decision rule for lateral transshipments in inventory systems, Manag. Sci. 49 (9) (2003) 1168–1179.
[97] S. Minner, E.A. Silver, D.J. Robb, An improved heuristic for deciding on emergency transshipments, European J. Oper. Res. 148 (2003) 384–400.
[98] G. Yang, R. Dekker, A.F. Gabor, S. Axsäter, Service parts inventory control with lateral transshipment and pipeline stock flexibility, Int. J. Prod. Econ. 142 (2013) 278–289.
[99] G. Hadley, T. Whitin, Analysis of Inventory System, Prentice-Hall, Englewood Cliffs, NJ, 1963.
[100] M.N. Jalil, Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics (Ph.D. thesis), Erasmus University Rotterdam, Rotterdam, The Netherlands, 2011.
[101] H.G.H. Tiemessen, M. Fleischmann, G.J. van Houtum, J.A.E.E. Van Nunen, E. Pratsini, Dynamic demand fulfillment in spare parts networks with multiple customer classes, European J. Oper. Res. 228 (2) (2013) 367–380.
[102] R. Hariharan, P.H. Zipkin, Customer-order information, leadtimes, and inventories, Manag. Sci. 41 (10) (1995) 1599–1607.
[103] R. Li, J.K. Ryan, A Bayesian inventory model using real-time condition monitoring information, Prod. Oper. Manage. 20 (5) (2011) 754–771.
[104] D. Louit, R. Pascual, D. Banjevic, A.K.S. Jardine, Condition-based spares ordering for critical components, Mech. Syst. Signal Process. 25 (2011) 1837–1848.
[105] X. Lin, R.J.I. Basten, A.A. Kranenburg, G.J. van Houtum, Condition based spare parts supply, BETA Working Paper 371, 2012.
[106] A.H. Elwany, N.Z. Gebraeel, Sensor-driven prognostic models for equipment replacement and spare parts inventory, IIE Trans. 40 (7) (2008) 629–639.
[107] L. Wang, J. Chu, W. Mao, A condition-based order-replacement policy for a single-unit system, Appl. Math. Model. 32 (11) (2008) 2274–2289.
[108] K. Wang, J. Chu, W. Mao, A condition-based replacement and spare provisioning policy for deteriorating systems with uncertain deterioration to failure, European J. Oper. Res. 194 (1) (2009) 184–205.
[109] M. Rausch, H. Liao, Joint production and spare part inventory control strategy driven by condition based maintenance, IEEE Trans. Reliab. 59 (3) (2010) 507–516.

[110] M.C. Van der Heijden, E.M. Alvarez, J.M.J. Schutten, Inventory reduction in spare part networks by selective throughput time reduction, Int. J. Prod. Econ. 143 (2013) 509–517.
[111] A. Díaz, M.C. Fu, Models for multi-echelon repairable item inventory systems with limited repair capacity, European J. Oper. Res. 97 (3) (1997) 480–492.
[112] A. Sleptchenko, M.C. Van der Heijden, A. Van Harten, Trade-off between inventory and repair capacity in spare part networks, J. Oper. Res. Soc. 54 (2003) 263–272.
[113] M.F. Candas, E. Kutanoglu, Benefits of considering inventory in service parts logistics network design problems with time-based service constraints, IIE Trans. 39 (2) (2007) 159–176.
[114] J.A. Rappold, B.D. Van Roo, Designing multi-echelon service parts networks with finite repair capacity, European J. Oper. Res. 199 (2009) 781–792.
[115] J.V. Jones, Integrated Logistics Support Handbook, third ed., McGraw-Hill, New York, NY, 2006.
[116] L.L. Barros, The optimization of repair decisions using life-cycle cost parameters, IMA J. Math. Appl. Bus. Industry 9 (1998) 403–413.
[117] H. Saranga, U. Dinesh Kumar, Optimization of aircraft maintenance/support infrastructure using genetic algorithms — level of repair analysis, Ann. Oper. Res. 143 (1) (2006) 91–106.
[118] R.J.I. Basten, J.M.J. Schutten, M.C. van der Heijden, An efficient model formulation for level of repair analysis, Ann. Oper. Res. 172 (1) (2009) 119–142.
[119] R.J.I. Basten, M.C. Van der Heijden, J.M.J. Schutten, A minimum cost flow model for level of repair analysis, Int. J. Prod. Econ. 133 (1) (2011) 233–242.
[120] R.J.I. Basten, M.C. Van der Heijden, J.M.J. Schutten, Practical extensions to a minimum cost flow model for level of repair analysis, European J. Oper. Res. 211 (2) (2011) 333–342.
[121] E.S. Brick, E. Uchoa, A facility location and installation or resources model for level of repair analysis, European J. Oper. Res. 192 (2) (2009) 479–486.
[122] P. Alfredsson, Optimization of multi-echelon repairable item inventory systems with simultaneous location of repair facilities, European J. Oper. Res. 99 (1997) 584–595.
[123] R.J.I. Basten, M.C. Van der Heijden, J.M.J. Schutten, E. Kutanoglu, An approximate approach for the joint problem of level of repair analysis and spare parts stocking, Ann. Oper. Res. (2014) in press.
[124] M.W.F.M. Draper, A.E.D. Suanet, Service parts logistics management, in: C. An, H. Fromm (Eds.), Supply Chain Management on Demand. Strategies, Technologies, Applications, Springer, Berlin, Germany, 2005, pp. 187–210 (Chapter 9).
[125] Y. Erke, Y.C. Ma, M.C. Booth, Method of determining inventory levels. Patent Application US 2003/0061126 A1, IBM Corporation, Endicott, NY, 2003.