Contents lists available at ScienceDirect

International Journal of Production Economics

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

Joint optimization of spare parts inventory and service engineers staffing with full backlogging

S. Rahimi-Ghahroodi[a,*], A. Al Hanbali[b], I.M.H. Vliegen[c], M.A. Cohen[d]

[a] Department Industrial Engineering and Business Information Systems, Faculty of Behavioural, Management and Social Sciences, University of Twente, P.O. Box 217, AE Enschede, 7500, the Netherlands

[b] Systems Engineering Department, College of Computer Science & Engineering, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia

[c] School of Industrial Engineering, Eindhoven University of Technology, 5600 MB, Eindhoven, the Netherlands

[d] The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

ARTICLE INFO

Keywords:
Service logistics
Spare parts inventory
Maintenance
Heuristics
Queueing

ABSTRACT

We consider the integrated planning of spare parts and service engineers that are needed for serving a group of systems. These systems are subject to different failure types, and for each failure, a service engineer with the necessary spare part has to be assigned to repair the system. The service provider follows a backlogging policy with part reservations. That is, a repair request is backlogged if one of the required resources is not immediately available upon demand. Moreover, a spare part is reserved if the requested spare part is in stock but no service engineer is immediately available. The spare parts are typically slow-movers and are managed according to a base-stock policy. The objective is to jointly determine the stock levels and the number of service engineers to minimize the total service costs subject to a constraint on the expected total waiting times of the repair calls. For the evaluation of a given setting, we present an exact method (computationally feasible for small problems) and an accurate approximation. For the joint optimization, we present a greedy heuristic that efficiently produces close-to-optimal results. We test how the heuristic performs compared to the optimal solution and the separate optimization of spare parts and service engineers in an extensive numerical study. In a case study with 93 types of spare parts, we show that the solution of the greedy algorithm is always within 2% of the optimal solution and is up to 20% better than a separated optimization approach encountered in practice.

1. Introduction

In the technically advanced business environment, system availability is generally crucial for a company's operations. An operational failure resulting in downtime is highly undesirable and can be very expensive. Different maintenance services such as inspections and preventive maintenance activities are executed with the goal to maximize the availability of these expensive systems. However, failures still happen and are often unavoidable. Therefore, in addition to preventive maintenance and services, repair actions are necessary. Because of the focus on the up-time of systems, often the policy "repair-by-replacement" is adopted, i.e. upon detection of what parts are malfunctioning, these parts are removed and replaced by functioning spare parts. In this case, spare parts and service engineers are the main resources for executing the repair actions and the availability of them has a major impact on overall system downtime. Most spare parts are expensive, and service providers typically keep the smallest possible number of them in stock while still satisfying agreements on availability with their customers. Furthermore, the repair of systems needs to be carried out by highly skilled and hence expensive service engineers. Hence, the optimal planning of spare parts inventory and service engineers staffing allows service providers to reduce the service costs considerably, but still meet the required high service levels.

With the increasing fierce competition in the after-sales market, many service providers are striving for a superb level of service offered to their customers and for maximizing their operational profit margin. Especially for advanced capital goods, the spare parts holding cost is among the major cost components hindering the maximization of the service provider profit margin. Therefore, the usage of advanced planning tools for spare parts as well as for service engineers is a common practice. Most service providers, however, decouple the above problem into two separate problems, i.e. one for spare parts inventory planning and one for service engineers staffing. One reason for this is that these resources are usually managed by different departments. The

[*] Corresponding author.
E-mail address: s.rahimighahroodi@utwente.nl (S. Rahimi-Ghahroodi).

https://doi.org/10.1016/j.ijpe.2019.02.007
Received 25 April 2018; Received in revised form 6 February 2019; Accepted 9 February 2019
Available online 15 February 2019

parts and service engineers are required jointly for a successful equipment repair. This makes the prompt satisfaction of a customer repair order mainly dependent on the availability of both the spare parts and the service engineers. Although it has been noted that the planning of parts and service engineers in an integrated way may result in a more efficient utilization of these two resources and in a better service delivery (AberdeenGroup, 2006), the integrated planning of these resources has received little attention so far in the literature. This is especially true for the case of a service provider following the standard “backlogging policy”, a widely used policy in practice. In this paper, we bridge this gap in the literature. We consider a set of equipment in an installed base. These systems are subject to random failures. Each failure (which results in a repair call) needs both a service engineer and a spare part to be available before the service provider can start to resolve it. We aim to quantify the performance gap between the current planning tool in the practice, i.e. separated planning, and the integrated planning of spare parts and service engineers for the full backlogging service policy.

The contribution of this paper is threefold. First, we analytically model the availability of the spare parts and the service engineers in an integrated way and evaluate the system performance. For the exact system evaluation, we analytically describe the model using a Markov chain. Second, since the exact evaluation method is computationally feasible only for small problems (small number of part types), we propose an accurate approximation method to evaluate large problems. Third, we study the integrated optimization problem to quantify the gain of jointly planning of spare parts and service engineers, compared to the results of the separate optimization. The objective of the integrated optimization problem is to determine the spare parts stock levels and the number of service engineers such that the total service cost is minimized subject to a service level constraint. In this model, the service level is made on the expected waiting time of the repair calls. A greedy heuristic algorithm is designed for solving the integrated optimization problem. To validate this algorithm, a lower bound for the optimal solution is proposed (for cases in which finding the optimal solution is not possible).

We discuss different approaches to the planning of spare parts inventory and service engineers that are typically carried out in practice. In these approaches, the spare parts inventory and the service engineers planning are done separately and the waiting times for both spare parts and service engineers are not taken into the account together in the planning stage. By doing so, either the service provider is not able to meet the service level requirement, or in cases where service level agreements are strictly enforced, it leads to over-staffing of the service engineers to make sure that the waiting time for them is negligible.

In a smarter way of separated optimization, the spare parts inventory and service engineers planning is achieved by splitting the maximum average waiting time between these two resources. In practice, the service provider decides on a splitting fraction before solving the problem and the chance of choosing the right fraction is very low. We show that, even by using the optimal splitting strategy, this approach to separated planning, results in a solution with up to 32% higher total cost than the optimal solution and 20% higher total cost than the solution of the integrated optimization heuristic that is presented in this paper (tested for problems with 10 types of spare parts).

In the remainder of this paper, we first describe the related literature in Section 2, followed by the model in Section 3. We then describe the Markov chain which is used to evaluate the performance of the system in Section 4. In Section 5, we investigate how to compute the total average waiting time efficiently. Then, we study the optimization problem in Section 6. A greedy algorithm, as well as three separated optimization methods, are described. Furthermore, a numerical validation of the approximation method and a comparison between the solutions found under all optimization approaches are given in the numerical study. In Section 7, we study a case using the optimization approaches discussed in Section 6. Finally, we draw conclusions and discuss managerial insights in Section 8.

2. Literature review

Although spare parts and service engineers have been studied extensively in the literature, most papers consider these two resources separately. We shortly review the literature on spare parts in Section 2.1 and service engineers in Section 2.2. There are only a few papers that study the joint planning problem. We discuss these papers in Section 2.3.

2.1. Spare parts

If we ignore the availability of service engineers, our problem is reduced to a spare parts inventory model. Spare parts management and optimization has been studied extensively in the literature starting with the seminal paper of Sherbrooke (1968). The literature of spare parts inventory management has been reviewed by Kennedy et al. (2002), Muckstadt (2004) and Sherbrooke (2006). For a more recent survey and review of spare parts management, we refer to Basten and van Houtum (2014), van Houtum and Kranenburg (2015) and Hu et al. (2018).

In the integrated planning of service engineers and spare parts, a request occupies simultaneously for two resources. However, in most spare parts models, it is assumed that only one part is needed to repair a failure. This feature makes our model theoretically different from spare parts management problems. A more general model considers the case where multiple failures occur simultaneously, each requesting a specific spare part. Only a few papers assume that multiple failures can happen. Some interesting studies on spare part models with multiple failures are van Jaarsveld et al. (2015), Cheung and Hausman (1995), Alt (1962), Miller (1971) and Schaefer (1983).

If we investigate general inventory models, the most similarity to our model is found in assemble to order systems (ATO). In this context, inventory models with demand for multiple items at the same time are considered (see Song and Zipkin, 2003; Bijvank, 2009; Atan et al., 2017, for an overview of research in this area). In both cases, there are simultaneous requests for multiple components/resources. Our problem is related to ATO for the case of multiple products with stochastic demand and replenishment lead time and base-stock inventory policy. Interesting references in this area are Wee and Dada (2010), van Jaarsveld and Scheller-Wolf (2015), Zhou and Chao (2012), Ko et al. (2011), Dayanik et al. (2003), Lu et al. (2005), Benjaafar and ElHafsi (2006), Zhao and Simchi-Levi (2006), ElHafsi et al. (2008), Lu (2008), Zhao (2009), Dogru et al. (2010), Lu et al. (2010) and Hoen et al. (2011). Note, in an ATO system, all resources (components) are consumable. However, in our problem, service engineers will again become available for possible future repair calls after finishing their service on a job.

2.2. Service engineers

A service provider also depends on other resources besides spare parts when providing service to its customers. The availability of service engineers is one of the main bottlenecks in ensuring that the service level agreements are met. Al Hanbali et al. (2015) consider human resources, where they focus on the assignment of a set of engineers to a group of customers with varying service level requirements. The authors analyze a non-preemptive $M/PH/c$ priority queue with various customer classes. The availability of service engineers, called manpower, is also studied in other research areas such as call centers and cross-training manpower planning (cf. Agnihothri and Mishra, 2004; Fırat and Hurkens, 2012) in which similar modeling structures are analyzed. An overview of these areas is discussed in Rahimi-Ghahroodi et al. (2017). Also, for a review of personnel scheduling and planning see van den Bergh et al. (2013). The service engineers planning problems have also been studied in simulation models, see, e.g. Dear and

Sherif (2000).

Besides spare parts and service engineers, the availability of service tools may sometimes also have a considerable influence on the total downtime of a system. There are a few studies that consider the service tools planning problem in a maintenance logistic system, see Vliegen and van Houtum (2009) and Vliegen (2009). In terms of service, tools are similar to service engineers (i.e. they are not consumed and reused). When a tool is needed, it is taken from the stock and will be returned after the repair is finished. Note, tools are usually ordered in sets and will be returned simultaneously (coupling in return). This is not the case for the integration of spare parts and service engineers as in this paper. The integration of service tools and spare parts planning is also considered in Vliegen (2009) using some simplifying assumptions.

2.3. Joint optimization problem

The joint optimization problem is linked to the repair kit problem. See Bijvank et al. (2010) for an extensive review on the repair kit problem. In the repair kit problem, there is only one service engineer who is carrying the kit of spare parts with him or herself and the replenishment of spare parts only occurs when the service engineer returns to the depot after a repair tour. In papers studying the repair kit problem, each tour of repairs is studied independently, which means that all spare parts (or tools) are restocked again directly after usage or at the end of each tour. Teunter (2006) studies the problem in which a repairman visits multiple locations before his repair kit is restocked. In this work, every tour is considered separately, which means that the replenishment lead times are not considered. Bijvank et al. (2010) extend the work of Teunter (2006) by introducing an exact formulation for the service level, instead of an approximation, while also considering other service policies. The models of Waller (1994) and Papadopoulos (1996) are special versions of the repair kit problem in which a revisit takes place when the repair job is not successful after the first visit.

Güllü and Köksalan (2013) study an optimization model for the kit-management problem with an exact evaluation of the system performance that is only tractable for small-scale problems. They propose a greedy heuristic procedure to find the base-stock levels that minimize the service costs, subject to a service level constraint. In this problem, items are stocked at a central location. Kits are composed of these items and sent to a customers site. From the kit, one item is used and the others are returned together to the central location after a certain holding time. The item that is used is replenished; the replenishment process is modeled as a finite capacity queue. If an item is not in stock, it is supplied via an emergency channel without any delay. As soon as a unit of that item becomes available again at the central location, it is returned to the emergency source.

Other papers use simulation as a methodology for performance analysis. Hertz et al. (2014) review the literature on simulation models in after-sales service logistics. There are some papers in spare parts management in which service engineers are considered, but they are always available. Caglar et al. (2004) mention the service engineer availability as a possible future research direction for a two-echelon spare parts inventory system. Tovia et al. (2010) study a service parts logistics system (SPLS) in which one service engineer is assigned to provide equipment service to a group of customers spread across a geographic region. The service engineer carries some spare parts according to a periodic review inventory policy. The service system is approximated with a modified $M/G/c$ queue with a head-of-the-line service discipline. Then, the system cost is approximated with a mathematical model, and a heuristic is described to obtain a close to optimal solution of the service engineer assignment given a fixed inventory policy. An integrated solution to the service engineer assignment and spare parts inventory policies is mentioned as a follow-up research.

Recently, the joint optimization problem of spare parts stock levels and service engineers staffing was studied in Rahimi-Ghahroodi et al. (2017). Similar to this paper, they investigate the integrated planning of spare parts and service engineers but for a different service policy, called the “partial backlogging policy”. They study a full emergency shipment in the case of spare parts stock-out and a backlogging policy for service engineers. Exact and approximate performance evaluation methods are proposed and the optimization problem is tested with a tailored optimization algorithm. They show that the integrated planning of the spare parts and service engineers can result in up to 27% cost savings compared to a separated optimization. In this paper, we focus on the common practice “full backlogging policy”. This leads to a completely different model for the performance evaluation of the system than the one in Rahimi-Ghahroodi et al. (2017). This new model requires different heuristic techniques that we specifically design. In both papers, the objective is to minimize the total service cost under a tight constraint on the average waiting time of a repair request. In the working paper by Rahimi-Ghahroodi et al. (2018), the result of this paper (full backlogging) is compared with Rahimi-Ghahroodi et al. (2017) (partial backlogging) and it is shown that none of these two service policies is always superior in terms of the total service cost given the same constraint on the average waiting time. Especially, for cases with an expensive emergency shipment cost and lower service level, the service provider is better off with the full backlogging policy. This intuitive result is an additional supporting argument to analyze the full backlogging policy. Lastly, Rahimi-Ghahroodi et al. (2018) use the result of this comparison in a game theoretic model of a two-echelon after-sales service network to guide the emergency supplier selecting a proper emergency shipment cost which is maximizing his profit and is acceptable for the service provider.

The joint optimization of the spare parts and the service engineers is also studied in Sleptchenko et al. (2018) where the service policy is to fully outsource the repair job when one of the resources, service engineers or spare parts, is not immediately available. In this service policy, they optimize the total costs without considering any constraint on waiting time or other service level agreements. Note, in some systems, outsourcing is not an option for the service provider or it is extremely expensive and the service provider has to rely only on his own resources to meet the service level agreement. In this situation, the service provider needs to follow the full backlogging service policy.

3. Model

We consider a single service region with multiple systems. A service provider is responsible for the maintenance of systems and has a team of service engineers and spare parts available for this. Different types of failure occur randomly in these systems. Each failure needs one unit of a specific spare part. The repair is done by replacement of a failed part with a ready-to-use spare part. A single stocking point is located in this region to supply various types of spare parts. Let $K = \{1,2, ...|K|\}$ denote the set of spare part types, and H_k be the cost of stocking part k for one month. The cost for having one service engineer available 24 h per day (note that this implies multiple service engineer shifts) is O Euros per month. When one of the systems breaks down, a demand occurs for a repair job for which both a specific spare part and a service engineer are needed. We assume that all engineers are qualified to execute the repairs for all of the parts. The repair job starts only when both resources are available. We assume that, whenever the necessary part is available for an arriving job, this part is assigned (reserved) to the job until a service engineer becomes available. At the same time, a replenishment for the reserved part is requested. This makes the service policy of each spare part type an FCFS policy. However, a service engineer is only assigned to the job once the needed spare part is available. The spare parts replenishment time is usually much larger than the engineer service time. Therefore, if we reserve a service engineer when the part is not available, he (or she) will be blocked for a long time (until the part becomes replenished), which is inefficient since another repair job

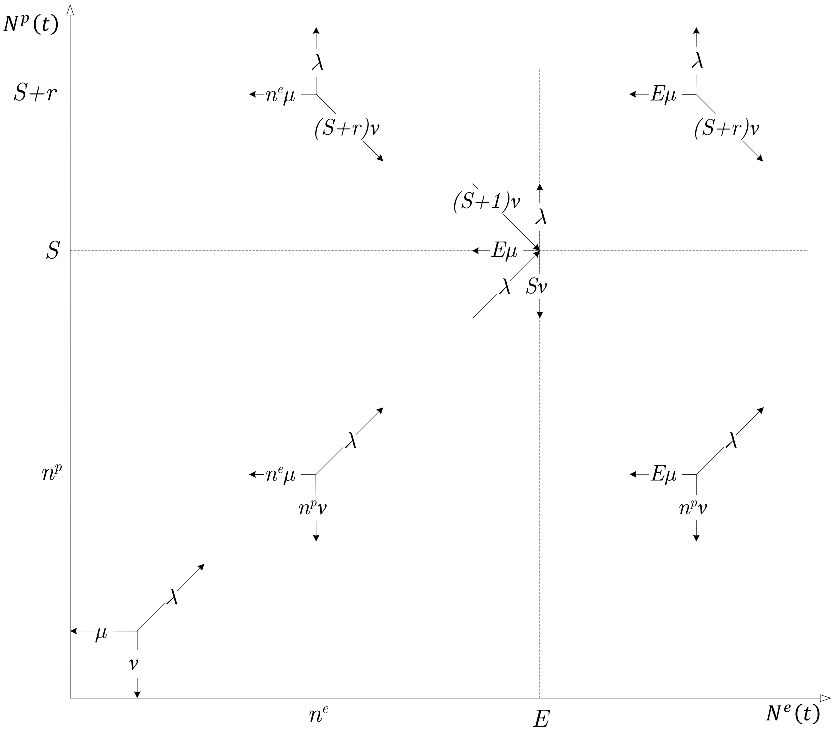

Fig. 1. Markov chain transitions diagram in a single part problem. $N^p(t)$ and $N^e(t)$ show the number of spare parts in the replenishment and the number of repair jobs waiting or being served in the service engineers queue, respectively.

can be executed during this time. Note that the service engineers might have some preventive maintenance activities as well, however, this work has a low priority compared to corrective maintenance (repair job). So whenever there is a request of repair job, they preempt their preventive jobs and go to satisfy the corrective one.

Repair jobs of type k, i.e. jobs that need a part type-k and a service engineer, arrive following a Poisson process with rate λ_k. The assumption of Poisson failures is standard in the spare parts literature and follows from the technical nature of the systems under consideration. For these systems, the mean time between failures is close-to exponential; see e.g., Munnik (2011) and Section 3.5.5 of Jardine and Tsang (2013). Furthermore, the service provider maintains multiple systems that all fail independently with low failure rate. Therefore, the failure rate observed by the service provider can be well approximated by a Poisson process. Whenever a demand for a repair job arrives, we assume that it is known immediately that a spare part and a service engineer are needed, and which spare part is required. So, contrary to Waller (1994) and Papadopoulos (1996), we assume that no inspection is needed, nor is there any doubt about the need for a part. We also assume that each repair requires only one spare part, which is standard for spare parts inventory models.

The inventory policy for spare part type k is a base-stock policy with base stock level S_k. This means that as soon as a demand occurs, a replenishment order is placed. This is a typical policy used for expensive slow-movers as in our case. The replenishment lead time for part k is exponentially distributed with mean $LT_k = 1/\nu_k$ months.

A total of E service engineers are available in the service region. The repair time is exponentially distributed with mean $RT = 1/\mu$ months, and is equal for all jobs independent of the part needed. The repair time includes traveling to the customer site and back to the service point, so, the service engineer is available again directly after the previous repair has finished.

The performance of the system is measured in terms of the waiting time, i.e., the time between the arrival of a repair job and the time the repair job has started. Waiting time is sufficient to determine the availability and up-time of the installed base of systems. Define $\mathbb{E}[WT]$ as the expected total waiting time. Of course, customers are mainly interested in the total time it takes to repair their system, which includes both the waiting time and the repair time. Since the repair time is assumed to be known and the target time to repair can be adjusted in the model, we focus on the waiting time.

The objective is to minimize the costs for spare parts and service engineers while satisfying the waiting time constraint agreed upon with the customers. Note, this waiting time constraint is over all the systems (not per system or machines). Let WT_{max} denote the maximum average waiting time allowed. The optimization problem thus becomes:

$$(\mathbf{P})\text{Min}\{\textstyle\sum_{k=1}^{K} H_k S_k + OE | \mathbb{E}[WT] < WT_{max}, \quad \mathbf{S} \in \mathbb{N}_0^K, E \in \mathbb{N}_1\},$$

where $\mathbf{S}$ is the vector of all spare parts stock levels. Note that cost of stocking a part is based on its target stocking level and not on the expected inventory. This is standard in spare parts management since the service provider pays for parts in the pipeline as well. The expected total waiting time is a non-linear function of decision variables, E and $\mathbf{S}$, which means $(\mathbf{P})$ is an integer non-linear optimization problem. Furthermore, as we show in Appendix A, the total expected waiting time is a non-convex function of spare parts stock levels. This makes $(\mathbf{P})$ a non-convex problem as well. Typically, for large problems of this type, one needs heuristics that yield suboptimal solutions. Nevertheless, before studying the optimization problem, we need to evaluate the model performance, or more precisely find the expected total average waiting time for given values of $\mathbf{S}$ and E. In Sections 4 and 5, we model the problem using Markov chains and queueing models to calculate the average total waiting time of a repair call, $\mathbb{E}[WT]$, in an efficient way.

4. Exact evaluation

In this section, we mathematically analyze this problem. To model the system we need to keep track of the number of repair jobs that are waiting or are being served by the service engineers queue N^e, and the number of outstanding replenishment orders for each spare part type k, N_k^p. It is well known that an inventory system state under a base-stock policy is fully described by the number of outstanding orders per type, and the same holds in our spare parts inventory system. Namely, if the number of type-k outstanding orders at time t, $N_k^p(t)$, is lower than the base stock level S_k we do have parts available. If $N_k^p(t) \geq S_k$, there are no parts in stock (although there may be parts reserved for repair jobs but still waiting for an engineer). If $N_k^p(t) > S_k$, there are jobs waiting for parts (in the inventory literature referred to as backorders). At the same time, we use N^e to record the state of the service engineers queue. Therefore, the whole system can be modeled by a $K + 1$ -dimensional Markov chain with state space $\{N^e(t), N_1^p(t), N_2^p(t), ..., N_K^p(t)\}$.

Fig. 1 shows the Markov chain representation of a situation with a single type of spare parts. The extension for multiple types of spare parts is straightforward. For more details and the description of the Markov chain transitions, we refer to Rahimi-Ghahroodi (2019), Section 4.3. The model described is infinite in all dimensions ($K + 1$). If we bound the model in K dimensions, the steady-state distribution can be solved using matrix-geometric analysis (see Rahimi-Ghahroodi, 2019, Appendix 4.A). Once we find the steady-state probabilities, we can calculate the expected repair calls waiting times for the spare parts and in the service engineers queue.

The matrix-geometric method is computationally tractable only for small problems ($K < 5$). In Section 5, we, therefore, investigate how we can find the total expected waiting time of repair calls in a more efficient way, such that it can be used for problems with a large number of spare parts types.

5. Expected total waiting time of the repair calls in the system

In this section, for a more efficient performance measurement, we analyze the spare parts inventory and the service engineers queue separately using queuing models. The total waiting time of the repair calls in the system is the sum of two parts: (1) W^p, the waiting time of repair calls for spare parts, if none is available on-hand; (2) W^e, the waiting time of repair calls for one of the service engineers to become available. As mentioned before, in the case that the requested spare part is available but there is no service engineer immediately available, the part is taken from the inventory and is reserved until a service engineer becomes available. This reservation makes the spare parts inventory independent of the service engineers queue. The same can be concluded by studying the marginal Markov chains, see Section 4. By looking at the equations of the marginal steady-state probabilities for each spare part type, we see that they are independent of the states of the service engineers queue as well as other spare part types inventory. However, this is not true for the marginal probabilities of the service engineers queue. The arrival process to the engineers queue depends on the spare parts stock levels. Therefore, the average waiting time of repair calls for spare parts can be determined independently, while for the exact evaluation of the service engineers queue, the impact of the spare parts stock levels on the arrival process should be considered.

5.1. Spare parts inventory

By definition, it is easy to show that for each type of spare part, the number of parts in the replenishment (outstanding orders) can be modeled as the number of jobs in an $M/M/\infty$ queue. This also can be derived from the marginal Markov chain of each spare part type. Therefore, the expected number of type-k repair calls waiting for spare parts, $\mathbb{E}[Q_k^p]$, $k = 1, ..., K$, is well known and given as follows, see, e.g. Sherbrooke (1968).

$$\mathbb{E}[Q_k^p] = \mathbb{E}[(Q_{M/M/\infty} - S_k)^+],$$
$$= \rho_k^p\left(1 - \sum_{i=0}^{S_k-1} \frac{(\rho_k^p)^i}{i!} e^{-\rho_k^p}\right) - S_k\left(1 - \sum_{i=0}^{S_k} \frac{(\rho_k^p)^i}{i!} e^{-\rho_k^p}\right), \tag{1}$$

where $\rho_k^p = \frac{\lambda_k}{\nu_k}$. The total expected waiting time (of repair calls) for spare parts gives

$$\mathbb{E}[W^p] = \frac{1}{\lambda} \sum_{k=1}^{K} \mathbb{E}[Q_k^p]. \tag{2}$$

It is easy to show that $\mathbb{E}[W^p]$ is a convex function in spare part stock level, see Sherbrooke (1968).

It is known that the steady state distribution of the number of busy servers in a $M/G/\infty$ queue is given in terms of only the arrival rate and the mean service time (see e.g., Tijms, 2003). It means Eq. (1) holds for the case where the spare parts replenishment lead time follows a non-exponential distribution as well. However, the output process from the spare parts inventory system, which will be the arrival process for the service engineers queue, does depend on the higher moments of the replenishment time distribution (see Rahimi-Ghahroodi, 2019, Appendix 4.F). In other words, a different replenishment time distribution even with the same mean will impact the average waiting time in the service engineers queue. Therefore, the total average waiting time in the system is sensitive to the replenishment time distribution.

5.2. Service engineers queue: aggregation approximation

The expectation of the waiting time in the service engineers queue is not easy to find. As described previously, the failure arrival streams to the spare parts inventory are Poisson processes. When a stock-out happens in the spare parts inventory, the repair call is backlogged until the requested spare part is replenished. The repair call joins the service engineers queue only once the part is available. This means, sometimes the repair call arrives at the service engineers queue with a delay (waiting for a part to be replenished in case of a stock-out). Therefore, the arrival process to the service engineers queue is a superposition of arrival streams for various types of spare parts that are not Poisson processes. The dependency of arrivals on spare parts stock inventory causes arrivals at the service engineers pool to be dependent on past arrivals. Therefore, the arrival streams to the service engineers queue are also non-renewal processes. Hence, to evaluate the service engineers queue separately from the spare parts inventory, we need to deal with a multi-server queue with superposition of non-renewal arrival streams, $\sum G/M/c$ queue with $c = E$, the number of service engineers.

There is no exact result in the literature for a $\sum G/M/c$ queue with non-renewal arrival streams. However, different approximations are studied and proposed for similar queues. Some problems have been studied in single failure spare parts inventory with lateral transshipment or commonality, in which they develop similar models for the system performance evaluation as we use in this paper, see Kranenburg and van Houtum (2007), Kranenburg and van Houtum (2009) and van Wijk et al. (2012). The lateral transshipment or commonality make the spare parts stocks dependent on each other. More precisely, the demand arrival to each spare parts inventory depends on the spare parts stocks in other locations, which makes the arrival streams non-Poisson processes. This is similar to the impact that spare parts inventory has on the repair calls arrival to the service engineers queue. In these papers, they mostly approximate these non-Poisson processes with Poisson processes. Other methods are also used such as the interrupted Poisson process (Kuczura, 1973). Other approximations such as mean value analysis and Laplace methods are studied in Rahimi-Ghahroodi et al. (2017), however, none of these methods gives satisfactory result for approximating the expected waiting time in the service engineers queue in this model. Therefore, we come up with a new method, the "aggregated approximation", based upon the exact average waiting time of scaled single item queues.

Aggregation approximation (AA method). For problems with one spare part type, the matrix-geometric method is fast enough (less than a couple of seconds) to calculate the exact average waiting time in the service engineers queue. For this approximation method, we use the result of the expected waiting time for single item problems as an approximation for the multi-item problems. Suppose there are K types of spare parts. For each spare part type k, $k = 1, \ldots, K$, we first determine an independent average service engineers waiting time, as follows. We change the arrival rate λ_k and the replenishment rate ν_k of this spare part to λ (total arrival rate) and $\lambda\frac{\nu_k}{\lambda_k}$ respectively, and keep the other parameters the same. This scaling keeps the offered load (workload) in the spare parts inventory $\rho_k = \lambda_k/\nu_k$ the same. Then, we find the average waiting time in the service engineers queue, W_k^s, by means of the matrix-geometric method. Next, we approximate total average waiting time in the service engineers queue by

$$\mathbb{E}[W^p] = \frac{\sum_k \lambda_k W_k^s}{\lambda}. \tag{3}$$

This method is efficient for any size of problems. We investigate the accuracy of this approximation in an extensive numerical experiment for instances with two to 50 types of spare parts, see Appendix A. Although there is no structural or bounding result to show the accuracy of the aggregation approximation, from the numerical result, we can conclude that AA works very well. As mentioned in Appendix A, this numerical test is designed such that it covers different situations, thus valid for testing the accuracy of the approximation. Compared to other approximation methods mentioned earlier, the AA method gives a much lower average and maximum error. Moreover, in contrast with the matrix-geometric method, there is no computational limitation in the use of the AA method. In addition, in instances where the AA gives negative errors (underestimate), the absolute difference between the exact (or simulation) results and the results of AA is negligible. Therefore, this approximation is appropriate to use for the optimization problem. By using the AA for optimizations, the chance of getting an infeasible solution (because of underestimation) is very low and if it happens we expect only very small deviations.

With the means of aggregate approximation, we are able to calculate the total expected waiting time of repair calls for any size of problems efficiently. In the next section, we study the procedures with which we can solve the optimization problem introduced in Section 3.

6. Optimal capacity decisions

In this section, we study the optimization problem of the integrated spare parts inventory and service engineers staffing. As introduced before in Section 3, the optimization problem (**P**) that we need to solve is as follows:

$$(\mathbf{P})\text{Min}\{TC(E, \mathbf{S}) | \mathbb{E}[WT] < WT_{max}, \quad \mathbf{S} \in \mathbb{N}_0^K, E \in \mathbb{N}_1\},$$

where,

$$TC(E, \mathbf{S}) = \sum_{k=1}^{K} H_k S_k + OE, \tag{4}$$

$$\mathbb{E}[WT] = \mathbb{E}[W^p] + \mathbb{E}[W^e]. \tag{5}$$

We want to find spare parts stock levels and a number of service engineers that minimize the total spare parts holding cost and the service engineer hiring cost, subject to the condition that the total repair calls expected waiting time is less than a promised level. As we notice in Appendix A, $\mathbb{E}[W^e]$ and therefore $\mathbb{E}[WT]$ is a non-convex function of spare parts stock level. This makes the problem (**P**) a non-convex integer non-linear optimization problem for which an efficient way of finding the exact optimal solution is not known. In Section 6.3, we propose an efficient greedy heuristic to solve this integrated optimization problem. To quantify the benefit of the integrated planning, however, we first discuss the separated optimization and the approaches that are used in practice in Sections 6.1 and 6.2. The accuracy of all methods is tested in a numerical experiment.

6.1. Optimizing the resources separately

Although it is clear that spare parts and service engineers have a joint effect on the service level, an often used method for optimizing the capacities of these resources is to split them. One reason for this is that in practice, these resources are usually managed in different departments. Here, we study three ways of splitting the problem that usually happen in practice:

- Ignoring the impact the other resource has, i.e., assuming the other resource has infinite capacity.
- Service engineers over-staffing to make sure the waiting time for them is negligible.
- Constraint splitting: Splitting the accepted total waiting time in an accepted waiting time for parts and an accepted waiting time for service engineers.

In the following, we discuss the first two approaches. The constraint splitting approach will be discussed in next section in more details.

In the first approach, companies often decouple these two resources capacity decisions and each decision is done by assuming that the other resource has infinite capacity, and the waiting time will be caused solely by the resource under consideration. The optimization subproblem faced by the department responsible for the spare parts then becomes:

$$(\mathbf{P_p})\text{Min}\left\{\sum_{k=1}^{K} H_k S_k | \mathbb{E}[W^p] < WT_{max}, \quad \mathbf{S} \in \mathbb{N}_0^K\right\}.$$

To solve the spare parts optimization subproblem, we use the known greedy heuristic in the spare parts inventory literature which is originally introduced by Sherbrooke (1968) and then extended later on. Note, this greedy heuristic does not necessarily give the optimal solution, but it performs well (see van Houtum and Kranenburg, 2015, p. 20–21).

The optimization subproblem for the service engineers is as follows:

$$(\mathbf{P_e})\text{Min}\{OE | \mathbb{E}[W^e] < WT_{max}, \quad E \in \mathbb{N}_1\}.$$

Since we separate the spare parts inventory and the service engineers queue, we have an $M/M/c$ queue with $c = E$ for the service engineers. This subproblem is easy to solve. The waiting time is decreasing in the number of service engineers. We start with the minimum number of service engineers that guarantees the queue stability. We increase the number service engineers one by one until the average waiting time becomes less than WT_{max}.

As is clear, this way of optimizing the resource capacities will lead to waiting times that might be much higher than the waiting time target. However, since we only allow for integer values of S and E, it is possible that a lower waiting time is achieved as well.

Often service providers have a very strict agreement on the service level with the customer and not satisfying the service level agreement will be very costly for them. Therefore, following the first approach is not an option for them. Still, in some industries, they only focus on the spare parts, and the stock levels are decided by assuming that the service engineers are always available for the repairs. Since the service level agreement with the customer is very strict, this way leads to overstaffing of service engineers to make sure the waiting time for service engineers is negligible. In this method, the spare parts stock levels can be found by solving the Problem ($\mathbf{P_p}$). To find the number of service engineers, we need to solve the following problem:

$$(\mathbf{P'_e})\text{Min}\{E | \mathbb{E}[W^e] < WT_{max} - \mathbb{E}[W^P(\mathbf{S_p})], \quad E \in \mathbb{N}_1\},$$

where $\mathbf{S_p}$ is the solution of the Problem ($\mathbf{P_p}$) and $\mathbb{E}[W^P(\mathbf{S_p})]$ the average waiting time in the spare parts inventory given this solution.

In the next section, we study the third way of decoupling the spare parts inventory and service engineers planning which is a smarter way of separation.

6.2. Constraint splitting

Another way of decoupling the decisions of resource capacities which seems to be more rational, is by splitting the maximum total waiting time in a waiting time for service engineers and a waiting time for parts. Let us define the splitting fraction, α, as the percentage of the total waiting time that can be caused by spare parts:

$$WT_{max}^{p} = \alpha WT_{max}, \tag{6}$$

$$WT_{max}^{e} = WT_{max} - WT_{max}^{p}. \tag{7}$$

Then, the optimization subproblems become as follows:

$(\mathbf{P_p})\mathrm{Min}\{\sum_{k=1}^{K} H_k S_k \| \mathbb{E}[W^p] < WT_{max}^{p}, \quad \mathbf{S} \in \mathbb{N}_0^K\}$.

$(\mathbf{P_e})\mathrm{Min}\{OE \| \mathbb{E}[W^e] < WT_{max}^{e}, \quad E \in \mathbb{N}_1\}$.

Each subproblem can be solved with the same procedure as presented for previous approaches. Note, similar to other separated approaches, the effect of spare parts stock levels on the service engineers queue is not considered in this method. Therefore, the average waiting time in the service engineers queue is calculated using $M/M/c$ queues result. It is worth noting that the second separated optimization approach discussed earlier (service engineers overstaffing) is an extreme case of the constraint splitting method where the splitting fraction, α, is equal to 1.

The performance of this method highly depends on how the total waiting time is split. To have a “smart” constraint splitting procedure, we need to find the best splitting strategy. One way to find a good value of α is to search in a limited number of possible values and choose the best among them. Depending on the number of possible values, we can not have the accuracy and efficiency together, especially for larger problems.

Suppose the number of service engineers, and thus the expected average waiting time in the service engineers queue are given. Then, the maximum average waiting for the spare parts is equal to

$$WT_{max}^{p} = WT_{max} - \mathbb{E}[W^e].$$

Thereafter, the spare parts stock level can be determined by solving the problem $(\mathbf{P_p})$. Hence, by having the number of service engineers, the splitting fraction can be easily calculated as follows:

$$\alpha = \frac{WT_{max} - \mathbb{E}[W^e]}{WT_{max}}.$$

This explanation suggests a smarter way to find the optimal maximum average waiting time splitting strategy:

Step 1. Define the range $[E^0, E^1]$ in which the optimal number of service engineers is included for sure. To find E^0, we assume that all the waiting time is caused by service engineers ($\alpha = 0$).

$$E^0 = \min\{E \| \mathbb{E}[W^e(E)] \leq WT_{max}\}$$

To find E^1, we assume that α is very close to one, e.g., 0.99. In real cases, a value for α higher than 0.99 does not happen in the optimal solution.

$$E^1 = \min\{E \| \mathbb{E}[W^e(E)] \leq 0.01\ WT_{max}\}$$

Step 2. For E in $[E^0, E^1]$:

$$WT_{max}^{p}(E) = WT_{max} - \mathbb{E}[W^e(E)]$$

Given $WT_{max}^{p}(E)$, $\mathbf{S}(E)$ is the solution of the problem $(\mathbf{P_p})$. Thus, the total cost is equal to

$$TC(E) = \sum_{k=1}^{K} H_k S_k(E) + OE.$$

Step 3. We keep the number of service engineers, E^*, which gives the lowest total cost value:

$$E^* = \mathrm{argmin}\{TC(E)\}.$$

Therefore, the equation below gives the optimal splitting fraction.

$$\alpha^* = \frac{WT_{max}^{p}(E^*)}{WT_{max}}.$$

As has been observed in some examples, the cost function associated with the optimal spare parts levels, given the number of service engineers E, is not always convex in E. Therefore, we should not stop the search when the total cost starts to increase. The search should be done for the whole range between E^0 and E^1. By searching in different numbers of service engineers, we actually only search for those values of α that are on the border of changing the number of service engineers. The best α value is for sure one of these values.

Note that, by finding the best splitting strategy, there is no guarantee that the constraint splitting method gives the optimal solution. First, for the spare parts subproblem, we use the standard greedy heuristic that may not be optimal. Second, we find the expected waiting time in the service engineers queue by assuming that it is an $M/M/c$ queue. However, in the real system, the arrival process is not Poisson and the exact expected waiting time may be smaller compared to an $M/M/c$ queue.

6.3. Greedy heuristic

In the integrated optimization problem $(\mathbf{P})$, since there is no analytical solution available, finding the optimal solution might be done by enumerating all possible solutions. However, we instead propose a greedy heuristic algorithm that is far more efficient. For the algorithm initial solution, we start with the minimal values of E and $\mathbf{S}$ as given below. We can find the minimal values by arguing that the average waiting time that is caused individually by each spare part inventory or by the service engineers must be at most equal to WT_{max}.

- S_k^0: minimal S_k, $k = 1, \ldots, K$. Assuming that the capacity for the service engineers and other spare parts is never limiting, we can determine the minimal number of each spare part stock level such that the average waiting time constraint is met.
- E^0: minimal E. The minimum number of engineers is easily determined by assuming that there is no waiting time for spare parts (the capacity of spare parts is not limited). In this situation, we know that the arrival process to the service engineers queue leads to a Poisson process. Therefore, to determine the minimal value for E, we assume that the service engineers queue is $M/M/c$ and find the minimum number of service engineers that is needed to meet the average waiting time constraint.

We start the algorithm with $(E^0, \mathbf{S}^0)$ where $\mathbf{S}^0$ is the vector of S_k^0 values for all $k = 1, \ldots, K$. This solution may be an infeasible solution for the integrated planning problem.

Step 1. Set $(E, \mathbf{S})$: $=(E^0, \mathbf{S}^0)$.
Step 2. Calculate the total average waiting time, $WT(E, \mathbf{S})$. If $WT(E, \mathbf{S}) \leq WT_{max}$, then $(E, \mathbf{S})$ is the optimal solution. Otherwise, go to the next step.
Step 3. Calculate Δ^e and for each type of spare part Δ_k^p, using formulas below. Δ gives the highest value.

$$\Delta^e = \frac{WT(E, \mathbf{S}) - WT(E+1, \mathbf{S})}{TC(E+1, \mathbf{S}) - TC(E, \mathbf{S})} \tag{8}$$

$$\Delta_k^p = \frac{WT(E, \mathbf{S}) - WT(E, \mathbf{S} + e_k)}{TC(E, \mathbf{S} + e_k) - TC(E, \mathbf{S})} \tag{9}$$

$$\Delta = max_k\{\Delta^e, \Delta_k^p\} \tag{10}$$

If Δ^e is the highest value, update $E: = E + 1$, otherwise update $S_k: = S_k + 1$ for k that Δ_k^p gives the highest value.

Calculate $WT(E, \mathbf{S})$. If $WT(E, \mathbf{S}) \leq WT_{max}$, go to the next step. Otherwise, repeat step 3.

Step 4. Perform a local search to decrease the total cost while the solution remains feasible. The last solution is the suboptimal solution.

Local search. Since we have an integer optimization problem, performing a local search may improve the solution considerably. We search in the following directions:

- Decrease the number of service engineers by one: Since the average waiting time in the service engineers queue may decrease by increasing the spare parts stock levels, we check whether by decreasing the number of service engineers by one, the solution remains feasible or not.
- Decrease the stock level of one of the spare parts by one: Decrease in one of the stock levels may keep the solution feasible, but it decreases the total cost.
- Decrease the number of service engineers by one and simultaneously increase the stock level of one of the spare parts by one: If the hiring cost of service engineers is higher than the holding cost of the spare part, we check whether by a change in this direction the solution remains feasible or not.

Service target overshoot. Another way of improving the greedy heuristic is to avoid an overshoot in the final solution. The greedy heuristic may give a solution in which the total average waiting time is much lower than the target level. It means we are putting more resources than needed. This is called an overshoot. One way to avoid this is to change Step 3 as follows. Instead of calculating the decreases in the total waiting time per one unit of cost, we calculate the decreases in the total waiting time towards the waiting time target level per unit of cost. Therefore, the Δ functions change as follows:

$$\Delta^e = \frac{WT(E, \mathbf{S}) - \max\{WT_{max}, WT(E+1, \mathbf{S})\}}{TC(E+1, \mathbf{S}) - TC(E, \mathbf{S})} \tag{11}$$

$$\Delta_k^p = \frac{WT(E, \mathbf{S}) - \max\{WT_{max}, WT(E, \mathbf{S} + e_k)\}}{TC(E, \mathbf{S} + e_k) - TC(E, \mathbf{S})} \tag{12}$$

We test the greedy algorithm numerically with and without this updated Δ functions. Using equations (11) and (12) instead of equations (8) and (9) improves the greedy algorithm in average, but in some cases, we may obtain much worse solutions. However, we can easily get the benefit of both techniques. In our following numerical studies, we solve each problem with both the original greedy algorithm and the updated one (no overshooting). At the end, we take the best solution of these two algorithms as the final solution of the greedy heuristic.

The greedy optimization methods that we study in this paper are similar to the approaches that are used in spare parts inventory optimization. For a review of spare parts inventory optimization, see van Houtum and Kranenburg (2015). There are some papers that consider local search along with greedy heuristics, see, e.g., Wong et al. (2005). The concept of service target overshoot is also investigated in a few papers, e.g. Sherbrooke (2006). However, to the best of our knowledge, the combination of the local search and overshooting has not been applied in the use of greedy algorithms in spare parts inventory.

6.4. Optimal solution and lower bounds

In this section, we are interested to find the optimal solution or a lower bound when finding the optimal solution is not possible, to show how much the results of the proposed optimization algorithms deviate from the optimal one. To find the optimal solution, we perform an enumeration. To determine a sufficiently small feasible region, we use the solution of the greedy algorithm. More precisely, we only search in solutions with total costs equal to or lower than the total cost in the greedy solution. Note, this enumeration is only possible for small problems.

We need the exact evaluation (given in Section 4) to know the enumeration leads to a guaranteed optimal solution. However, the use of exact evaluation in the optimization is only tractable for instances up to three types of spare parts. For larger instances, we propose to use an underestimate approximation (lower bound) of the exact average waiting time in the service engineers queue as it is described in Rahimi-Ghahroodi (2019), Appendix 4.D. In this case, the enumeration always results in a lower bound of the optimal solution.

For large problems ($K > 10$), enumeration becomes almost impossible, even when using the lower bound procedure. We propose another method using the greedy procedure to find a lower bound for the optimal solution without enumeration. This method is introduced in Rahimi-Ghahroodi (2019), Appendix 4.E. In our following numerical experiments, we use this (greedy) lower bound in our large instances (where enumeration in search for the optimal solution is not possible) to test the performance of the proposed optimization algorithms.

6.5. Optimization algorithms performance

In this section, we investigate how the greedy algorithm performs compared to the separated optimizing of the resources and to the optimal solution or the lower bounds defined in Section 6.4. We use the same parameter settings as in Table A 3. We solve 6000, 2000, and 100 instances with two, five, and ten types of spare parts, respectively. In all instances, the hiring cost of service engineers is fixed to 100 €/day and the holding cost of spare parts varies from 10 to 400 €/day. Note, these cost factors are scaled and not necessarily realistic. However, they are chosen such that all ratio possibilities of service engineers hiring cost to spare parts holding costs ($[0.1, 4]$) are covered in the numerical experiment. Therefore, for each algorithm, we expect the same performances in real cases as in this numerical experiment.

In Sections 6.1 and 6.2, we discussed three ways for the optimization of spare parts and service engineers separately. In the first way, we optimize each resource separately by assuming that the other resource capacity is infinite, which is an often common practice in industry. As we discussed before, this way may end up in an infeasible solution that does not meet the waiting time constraint. For instances with two types of spare parts in 61% of cases, five types of spare parts in 72% of cases, and ten types of spare parts in 76% of cases we get infeasible solutions using this separated optimization approach.

Table 1 shows the average and the maximum relative errors of the total cost of the greedy algorithm and the “constraint splitting” solutions compared to the optimal solutions ($K = 2$) or the lower bounds ($K = 5$,10). Note, in the constraint splitting method, we search for the best splitting strategy by searching in the possible values of the number of service engineers as explained in Section 6.2.

In Table 1, the errors that are shown for the instances with 5 and 10 types of spare parts are upper bounds of the real errors since we use a lower bound for the average waiting time in the enumeration of the optimal solution. The real error may be much lower than the presented errors in Table 1. Moreover, in the table, the expression “Optimal Solution (%)” for instances with 5 and 10 types of spare parts shows the percentage of instances in which the greedy or constraint splitting algorithm gives the same solution as the solution associated with the lower bound. Therefore, we expect that in a higher percentage of instances the optimal solution is obtained using greedy and constraint splitting algorithms.

The maximum and average error that is given in Table 1 for the constraint splitting method is done by using the best α value (search in the number of service engineers). However, the constraint splitting never outperforms the integrated optimization with the greedy

Table 1
Maximum and average total cost percentage error for the greedy and the constraint splitting algorithms in the cases with two (6000 instances), five (2000 instances) and 10 types (100 instances) of spare parts. The percentage number of instances in which each algorithm gives the optimal solution is given. Runtime speed ratio shows in average how much faster each algorithm solves the problem compared to the optimal optimization (enumeration).

K		Greedy	Constraint Splitting
2	Average error (%)	0.085	3.07
	Maximum error (%)	23.26	90.34
	Optimal Solution (%)	98.35	64.28
	Runtime speed ratio	15	35
5	Average error (%)	0.99	4.44
	Maximum error (%)	16.88	38.27
	Optimal Solution (%)	67.90^{*}	27.48^{*}
	Runtime speed ratio	80	600
10	Average error (%)	0.759	4.89
	Maximum error (%)	5.05	31.45
	Optimal Solution (%)	$60.\ 0^{*}$	$14.\ 0^{*}$
	Runtime speed ratio	5700	140000

heuristic. First, the effect of spare parts stock levels on the average waiting time in the service engineers queue is not considered in the constraint splitting method ($M/M/c$ assumption). Second, the constraint splitting method cannot avoid overshooting like we managed to do with the proposed greedy heuristic. Moreover, in practice, the splitting fraction (α) of the constraint splitting method is usually chosen by the rule of thumb and most of the time it is far from its optimal value.

As noted earlier, the second separated planning method discussed in Section 6.1 (service engineers over-staffing) is actually an extreme case of the constraint splitting algorithm ($\alpha = 1$). Therefore, we do not include it in this performance comparison. However, to show that how a common procedure in practice can go far from the optimal strategy, we compare it with the greedy heuristic and the lower bound in a case study in Section 7.

These numerical results show that although the constraint splitting algorithm is faster than the greedy algorithm, its average error is considerably higher. In addition, the chance of finding the optimal solution is much higher using the greedy algorithm. We cannot draw a specific conclusion under which condition each of these optimization algorithms performs best or shows a higher error. However, using the greedy heuristic seems to be the best option for the optimization in terms of average and maximum deviation from the optimal solution.

For real-life problems with many types of spare parts, the efficiency becomes more critical and finding the optimal solution with enumeration is almost impossible. In our instances with 10 types of spare parts, the average runtime of the greedy algorithm (on a Core i5 computer) is only 8 s. The runtime should not be more than a couple of minutes for real size problems. For larger problems, obtaining the optimal solution using the greedy algorithm is less likely. However, we expect smaller errors in large problems, since the steps in the greedy procedures will be smaller compared to the total cost value (compare maximum error for instances with 2, 5 and 10 types of spare parts).

We can conclude that for instances with sufficiently many types of spare parts, the greedy algorithm will generate reasonable suboptimal solutions for Problem (**P**) with low computational effort. If more accuracy is needed, the optimal solution can be found using the enumeration and with much more computations. To make it faster, the greedy algorithm result can be used as an upper bound in the enumeration (as we did in our instances). Still, as can be seen in Table 1, the greedy algorithm on average is 5700 times faster than optimal enumeration for problems with 10 types of spare parts. This runtime difference will be much more for larger problems. As another advantage, the solution generated by the greedy algorithm, will be rather robust against the changes in parameters value in contrast with the optimal solution, see, e.g. van Houtum and Kranenburg (2015).

7. Case study

In this section, we perform a numerical study on a real size problem considered in Rahimi-Ghahroodi et al. (2017) as a case study. In this problem, an OEM of advanced assets used in the defense sector is responsible for service of a number systems which are operated by his customers. There are 93 different spare parts needed for corrective maintenance of these systems and there is a team of service engineers responsible for the repair.

Most of the parameters are based on data obtained from the company involved in this case study, however, the provided data is limited. First of all, there is not enough information regarding the service times and we have to estimate the service rates. We assume that these rates are the same for all types of spare parts. All these spare parts regard a single system and are at the same level in the product configuration tree. Therefore, the replacements of spare parts have a similar complexity, so the service rates are roughly the same for all types of repair jobs. Moreover, the company's data only includes the average value for different parameters. Therefore, interpreting the distributions from the data is not possible. For the failures, the mean time to failure for each part is provided. The failures data are associated to more than 200 identical set of assets. Therefore, we believe the installed base size is large enough that the assumption of having Poisson failure arrivals is justified (based on PalmKhintchine theorem, cf. Heyman and Sobel, 2003). In addition, these data are based on systems in their mature phase and the assets are not yet suffering from wear out. Therefore, there is no big increase in failure rates over time. Last, the objective of this case study is to test the accuracy of approximation method and the optimization algorithms for a rather large size problem (93 different parts), not as a tool to validate the assumptions. Therefore, the replenishment time and the service time are assumed to be exponentially distributed as a simplified assumption. The holding cost of a spare part per year is equal to 20% of the new part price and the service engineers hiring cost per year is equal to €200000 (four shifts, hence the employer's cost of an individual engineer are €50000 per year). The total failure arrival rate is about 250 failures per year. The average service time of the service engineers is equal to 10 h and the average replenishment time of the spare parts are in the range of 2100–8600 h.

We solve this problem with the greedy heuristic (integrated optimization) and constraint splitting method (smart separated optimization) for different target service levels (WT_{max}). The results are summarized in Table 2. The current practice of the company is to hire enough number of service engineers to make sure that there is no queueing for them (second separation method in Section 6.1). We solve the problem with this strategy as well and the result can be seen in the column “Practice”. To check the results, we compare the solutions with the greedy lower bound (see Rahimi-Ghahroodi, 2019, Appendix 4.E). For each algorithm, Table 2 shows the (percentual) deviations (error) of

Table 2
The greedy algorithm, the constraint splitting and the practice solutions' deviations from the lower bound of the optimal solution for cases with different maximum average waiting times (service levels). QT_{max} shows the maximum accepted average number of repair calls waiting in the system.

QT_{max}	$WT_{max}(hr.)$	Greedy algorithm error (%)	Constraint Splitting error (%)	Practice (%)
0.025	0.3	0.22	1.05	12.39
0.075	0.9	0.75	1.73	15.64
0.15	1.8	0.07	0.75	14.03
0.25	3	0.18	0.36	15.58
0.375	4.5	1.95	2.35	20.91
0.5	6	0.26	1.44	17.68
0.75	9	0.29	1.58	16.23
1	12	0.06	2.34	21.86

the (sub)optimal total cost from the lower bound value. Note, these errors are an upper bound on the real error values. To have a better understanding of the average waiting time, QT_{max} is also given in the table. It shows the maximum average number of repair calls waiting in the system which can be easily obtained from the WT_{max} (Little's law).

As can be seen in Table 2, the greedy solution is always within 2% of the optimal solution. For applications in practice, this is a reasonable and sufficient result. In comparison with constraint splitting optimization, the greedy algorithm yields always better results. The difference in this case study is rather small, but as it is shown in Table 1, the constraint splitting method may yield a solution with a large deviation from the optimal solution. The gap between the total cost of the current practice of the company and the solution of the greedy algorithm is considerable. The high cost of the practice strategy mostly comes from the cost of hiring a lot of service engineers. However, as the solution of the integrated planning suggests, by stocking more spare parts, the company is able to hire a fewer number of service engineers and allows queueing for them, but still, meet the service level agreement. Note, the total service cost in maintenance logistics is huge in real problems, and only 1% saving may be equivalent to a few million Euros per year.

For a more extensive analysis, we refer to a sensitivity analysis presented in Rahimi-Ghahroodi (2019), Section 5.2. The most important observation in this sensitivity analysis is regarding the optimal splitting fraction. Respect to change of parameters, a fluctuating and unpredictable behavior of the optimal splitting fraction is observed in all cases. This suggests that it is hard to derive a “rule of thumb” for a good value of the average waiting time splitting fraction before solving the integrated optimization that guarantees a close to optimal splitting strategy. This emphasizes more on the need for integrated planning in this problem. Even by using a smart separated optimization technique like constraint splitting method, there is no guarantee to achieve a good suboptimal solution, especially if the splitting fraction is chosen by a rule of thumb.

8. Conclusions

In this paper, we study an integrated planning of spare parts inventory and service engineers under a full backlogging policy. In case the requested spare part is available but there is no available service engineer immediately, the part is reserved and the request is backlogged until a service engineer becomes available. However, when the requested part is not available, a service engineer is not reserved and one is called for duty once the part is replenished. This policy makes the availability of service engineers dependent on the spare parts stock levels. We have developed an exact method for the performance evaluation of the system for a given policy using a matrix-geometric analysis. This method is efficient only for small problems. We calculate the exact average waiting time of the repair calls in the spare parts inventory using queuing models. We design the aggregated approximation for the average waiting time in the service engineers queue for larger problems where using the matrix-geometric method computationally is not tractable. In a numerical study, we show that the aggregation approximation gives satisfactory results for the average waiting time in the service engineers queue.

To optimize the spare parts inventory and service engineers queue in an integrated way, we propose the greedy heuristic. To quantify the benefit of the integrated optimization, we discuss the separated optimization approaches that are often used in practice. Two common practice separated optimization methods are investigated. We show that these algorithms either mostly give infeasible solutions or solutions far from the optimal strategy. A smarter separated optimization method, constraint splitting, is studied. In this method, the maximum total average waiting time is split for spare parts and service engineers. In a numerical study, we compare the result of the constraint splitting method and the greedy heuristic with the optimal solution or its lower bound. Although the constraint splitting method is faster, it never outperforms the greedy algorithm. For real size problems, the greedy algorithm generates reasonable suboptimal solutions with low computational effort.

We test the proposed optimization algorithm in a rather large size case study problem. We show that the solution of the greedy algorithm is always within 2% of the optimal solution and up to 20% better than the current practice approach. To get more managerial insight, we check the sensitivity of the problem to different optimization parameters. We show the fluctuating behavior of the optimal splitting fraction (α) versus different parameters which indicates that a rule of thumb for splitting the maximum average waiting time between spare parts and service engineers is not intuitive.

There is another service policy in between the ones studied in this paper and Rahimi-Ghahroodi et al. (2017), that seems to be different, but results in a similar model structure. When a spare part is not available, only the requested part is ordered via a fast emergency shipment. The request for the service engineers is then sent when the part has arrived at the parts storage location. Considering this policy, the system can be modeled with a similar quasi birth-death Markov chain and can be solved with the matrix-geometric method that is used in this paper (with some small modification). However, the proposed approximation methods for the average waiting time in the service engineers queue should be tested to see whether it also gives accurate results for this scenario.

By quantifying the benefit of integrated planning of spare parts and service engineers under a full backlogging service policy, we show that there is a considerable potential cost saving in the joint optimization of these resources. By separating the planning of these two resources, we are neglecting the impact of these resources on the availability of each other. Moreover, meeting the service level agreement cost efficiently is not possible without an integrated planning of resources. In addition, by comparing the result of this paper with Rahimi-Ghahroodi et al. (2017), we are able to see in what conditions it is beneficial for the service provider to choose the full backlogging policy and when the partial backlogging policy gives him lower total service cost; see the working paper Rahimi-Ghahroodi et al. (2018).

In this study, we assume the service time is equal to all types of repair jobs and is independent of the part needed. However, the Markov chain considered in this paper can handle having different service rate for different repair jobs. For this case, the representation of the Markov chain will be more complex. This is because in this case one needs to keep track of the type of job under repair. Therefore, it is not enough to know the total number of jobs waiting in the queue but also the type of jobs should added to the state of the Markov. Although this Markov chain will have larger state space, the matrix-geometric method can solve this new well-structured Markov chain.

It is well known that for a $G/G/c$ queue the average waiting time is sensitive to the service time distribution (e.g., Tijms, 2003). It means, if we assume non-exponential service time in this problem (even with the same mean), the provided results for the average waiting time in the service engineers queue will change. This is the same for the replenishment time distribution. Although the average waiting time in the spare parts inventory is not sensitive to the second and higher moments of the replenishment time distribution, having a non-exponential replenishment time will impact the arrival process to the service engineers queue. This makes the total expected waiting time also sensitive to the spare parts replenishment time distribution. Hence, it is not possible to simply relax the assumption of having exponential service and replenishment time without considering its effect on the total expected average waiting time. Nevertheless, the Markov chain introduced as the representation of the joint model and the matrix-geometric method used to analyze this Markov chain are flexible to handle the rich class of phase-type distributions. Of course, in that case, the Markov chain will be more complex and it will have more states and transitions. For non-phase-type distributions, we lose the tractability of the Markov analysis in this model. Therefore, extending the methodology of this paper to cover these cases is not straightforward and one may need other techniques such as simulation to analyze the model under these distributions.

In this paper, we consider a single echelon, single indenture model with homogeneous skilled service engineers. To study real problems, we may need to extend the presented model in different aspects. The model can be easily extended to non-homogeneous skilled service engineers. In the case of cross-trained engineers, more analysis and modification in the model is needed. The service constraint in this model is a waiting time constraint that is averaged over all repair call types. One may impose the service constraint per repair calls (per spare part type) or a constraint on the fraction of the repairs that should be done within a time window. We expect that the optimization procedures that are proposed in this paper would work for these service constraints as well. However, for the evaluation of these constraint, other methods are needed.

Acknowledgment:

This publication was made possible by the NPRP award [NPRP 7-308-2-128] from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Appendix A. Accuracy of aggregation approximation and $M/M/c$ result

In this section, we show that how the aggregation approximation method performs compared to the exact result. Before testing the aggregated approximation, we show what result we get if we just assume that the arrival process to the service engineers queue is a Poisson process. By this assumption, we can use the result of $M/M/c$ queues. In this way, we actually neglect the effect of the spare parts inventory on the arrival process at the service engineers queue. Therefore, the expected waiting time in the service engineers queue is independent of spare parts stock levels. When the spare parts stock levels are zero or very high, the exact expected waiting time is the same as the result of the $M/M/c$ queue analysis. However, for some values of stock levels, the expected waiting time is considerably lower than the $M/M/c$ waiting time. The squared coefficient of variation of the arrival process to the service engineers queue is always less than 1 (see Rahimi-Ghahroodi, 2019, Appendix 4.F). This means that the arrival process to the service engineers queue is less variable than a Poisson process. Thus, we expect a lower average waiting time in the service engineers queue than the average waiting time of an $M/M/c$ queue.

For a simple example with one type of spare part, Figure A 2 shows how the exact average waiting time in the service engineers queue deviates from the $M/M/c$ average waiting time for different values of replenishment rate and spare part stock level. Numerically, we observe that for a fixed value of the replenishment rate, the minimum average waiting time happens for a stock level with a value around λ/ν, i.e., when $\lambda/S\nu$ is around one. As a side remark, we can see that.

Remark 1. The exact expected waiting time is not a convex function respects to the spare parts stock levels.

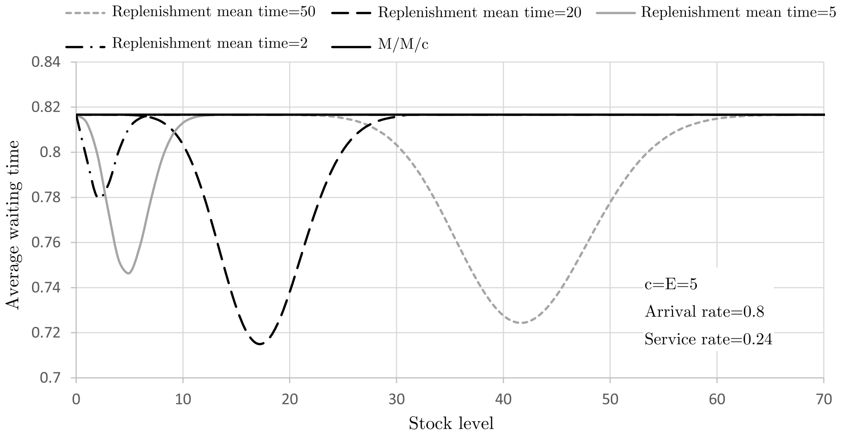

Fig. A.2. Average waiting time in the service engineers queue compared to the $M/M/c$ average waiting time in an example with one type of spare parts. This chart shows the deviation for different values of replenishment rate and stock level, which reaches its maximum around λ/ν.

In summary, although the $M/M/c$ result is very easy to use and reliable (it always overestimates the average waiting time in the service engineers queue), it does not give satisfactory accuracy for approximating the average waiting time in the service engineers queue. In the following, we examine the accuracy of the aggregation approximation in a numerical study.

Table A.3 shows the maximum and average error of the approximate expected waiting time using the aggregation approximation (AA) compared to the exact expected waiting time for a set of instances with two to fifty types of spare parts. To calculate the exact expected waiting time, we use the matrix-geometric method for problems with two types of spare parts ($K = 2$), and discrete-event simulation for problems with $K \geq 5$. We simulate the system using AnyLogic software package. To ensure that the results of the simulation are reliable, a proper warm-up period, run length and number of runs are applied. In our experiment, we have the following parameter setting. The total arrival rate is fixed to 1 per day. The other rates are chosen such that the parameter setting covers all the values of queues workload that can happen in practice. In this case, this numerical experiment is valid for testing the accuracy of the approximation. The chosen ranges for the workload of the service engineers queue and the spare parts inventory are [0.05, 1) and [0.04, 1) respectively. Therefore, the replenishment rate varies from 0.1 to 1.5 and the service rate from 0.2 to 4 per day. The spare parts stock levels are chosen in the range of $1 - 15$ units. In all defined ranges, a limited number of values are considered and instances are generated using a full factorial design. The number of service engineers is five in all instances.

Table A.3
Total expected waiting time error using the aggregation approximation in comparison with the exact evaluation for problems with different types of spare parts.

Number of part types			
(Number of instances)	Average error (%)	Max positive error (%)	Max negative error (%)
K = 2 (1350)	0.35	1.53	−0.18
K = 5 (185)	0.31	1.73	−0.11
K = 20 (56)	0.16	1.29	−0.09
K = 50 (28)	0.14	1.35	−0.13

References

AberdeenGroup, 2006. The Convergence of People and Parts in the Service Chain: Benchmarking the Alignment of Service Labor and Inventory Management. AberdeenGroup, Boston.

Agnihothri, S.R., Mishra, A.K., 2004. Cross-training decisions in field services with three job types and server–job mismatch. Decis. Sci. J. 35 (2), 239–257.

Al Hanbali, A., Alvarez, E., van der Heijden, M., 2015. Approximations for the waiting-time distribution in an m/ph/c priority queue. Spectrum 37 (2), 529–552.

Alt, F.L., 1962. Safety levels in military inventory management. Oper. Res. 10 (6), 786–794.

Atan, Z., Ahmadi, T., Stegehuis, C., de Kok, T., Adan, I., 2017. Assemble-to-order systems: a review. Eur. J. Oper. Res. 261 (3), 866–879.

Basten, R.J.I., van Houtum, G.J., 2014. System-oriented inventory models for spare parts. Surv. Oper. Res. Manag. Sci. 19 (1), 34–55.

Benjaafar, S., ElHafsi, M., 2006. Production and inventory control of a single product assemble-to-order system with multiple customer classes. Manag. Sci. 52 (12), 1896–1912.

Bijvank, M., 2009. Service Inventory Management: Solution Techniques for Inventory Systems without Backorders. Ph.D. thesis. Vrije Universiteit Amsterdam.

Bijvank, M., Koole, G., Vis, I.F., 2010. Optimising a general repair kit problem with a service constraint. Eur. J. Oper. Res. 204 (1), 76–85.

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

Cheung, K.L., Hausman, W.H., 1995. Multiple failures in a multi-item spares inventory model. IIE Trans. 27 (2), 171–180.

Dayanik, S., Song, J.-S., Xu, S.H., 2003. The effectiveness of several performance bounds for capacitated production, partial-order-service, assemble-to-order systems. Manuf. Serv. Oper. Manag. 5 (3), 230–251.

Dear, R.G., Sherif, J.S., 2000. Using simulation to evaluate resource utilization strategies. Simulation 74 (2), 75–83.

Dogru, M.K., Reiman, M.I., Wang, Q., 2010. A stochastic programming based inventory policy for assemble-to-order systems with application to the w model. Oper. Res. 58 (4-part-1), 849–864.

ElHafsi, M., Camus, H., Craye, E., 2008. Optimal control of a nested-multiple-product assemble-to-order system. Int. J. Prod. Res. 46 (19), 5367–5392.

Fırat, M., Hurkens, C., 2012. An improved mip-based approach for a multi-skill workforce scheduling problem. J. Sched. 15 (3), 363–380.

Güllü, R., Köksalan, M., 2013. A model for performance evaluation and stock optimization in a kit management problem. Int. J. Prod. Econ. 143 (2), 527–535.

Hertz, P., Cavalieri, S., Finke, G.R., Duchi, A., Schönsleben, P., 2014. A simulation-based decision support system for industrial field service network planning. Simulation 90 (1), 69–84.

Heyman, D.P., Sobel, M.J., 2003. Stochastic Models in Operations Research: Stochastic Optimization, vol. 2 Courier Corporation.

Hoen, K.M., Güllü, R., van Houtum, G.J., Vliegen, I.M.H., 2011. A simple and accurate approximation for the order fill rates in lost-sales assemble-to-order systems. Int. J. Prod. Econ. 133 (1), 95–104.

Hu, Q., Boylan, J.E., Chen, H., Labib, A., 2018. OR in spare parts management: a review. Eur. J. Oper. Res. 266 (2), 395–414.

Jardine, A.K., Tsang, A.H., 2013. Maintenance, Replacement, and Reliability: Theory and Applications. CRC press.

Kennedy, W., Patterson, J.W., Fredendall, L.D., 2002. An overview of recent literature on spare parts inventories. Int. J. Prod. Econ. 76 (2), 201–215.

Ko, S.-S., Choi, J.Y., Seo, D.-W., 2011. Approximations of lead-time distributions in an assemble-to-order system under a base-stock policy. Comput. Oper. Res. 38 (2), 582–590.

Kranenburg, A.A., van Houtum, G.J., 2007. Effect of commonality on spare parts provisioning costs for capital goods. Int. J. Prod. Econ. 108 (1), 221–227.

Kranenburg, A.A., van Houtum, G.J., 2009. A new partial pooling structure for spare parts networks. Eur. J. Oper. Res. 199 (3), 908–921.

Kuczura, A., 1973. The interrupted Poisson process as an overflow process. Bell Syst. Tech. J. 52 (3), 437–448.

Lu, Y., 2008. Performance analysis for assemble-to-order systems with general renewal arrivals and random batch demands. Eur. J. Oper. Res. 185 (2), 635–647.

Lu, Y., Song, J.-S., Yao, D.D., 2005. Backorder minimization in multiproduct assemble-to-order systems. IIE Trans. 37 (8), 763–774.

Lu, Y., Song, J.-S., Zhao, Y., 2010. No-holdback allocation rules for continuous-time assemble-to-order systems. Oper. Res. 58 (3), 691–705.

Miller, B.L., 1971. A multi-item inventory model with joint backorder criterion. Oper. Res. 19 (6), 1467–1476.

Muckstadt, J.A., 2004. Analysis and Algorithms for Service Parts Supply Chains. Springer Science & Business Media.

Munnik, M., 2011. Research on the Management of Service Level Agreements at Océ (Ph.D. thesis). .

Papadopoulos, H., 1996. A field service support system using a queueing network model and the priority mva algorithm. Omega 24 (2), 195–203.

Rahimi-Ghahroodi, S., 2019. Integration and Coordination in After-Sales Service Logistics. Ph.D. thesis. University of Twente, Netherlands.

Rahimi-Ghahroodi, S., Al Hanbali, A., Zijm, W.H.M., Timmer, J., 2018. Emergency Supply Contracts for a Service Provider with Limited Local Resources (Working paper). .

Rahimi-Ghahroodi, S., Al Hanbali, A., Zijm, W.H.M., van Ommeren, J.K.W., Sleptchenko, A., 2017. Integrated planning of spare parts and service engineers with partial backlogging. Spectrum 1–38.

Schaefer, M.K., 1983. A multi-item maintenance center inventory model for low-demand reparable items. Manag. Sci. 29 (9), 1062–1068.

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

Sherbrooke, C.C., 2006. Optimal Inventory Modeling of Systems: Multi-Echelon Techniques, vol. 72 Springer Science & Business Media.

Sleptchenko, A., Al Hanbali, A., Zijm, H., 2018. Joint planning of service engineers and spare parts. Eur. J. Oper. Res. 271 (1), 97–108.

Song, J.-S., Zipkin, P., 2003. Supply chain operations: assemble-to-order systems. Handb. Oper. Res. Manag. Sci. 11, 561–596.

Teunter, R.H., 2006. The multiple-job repair kit problem. Eur. J. Oper. Res. 175 (2), 1103–1116.

Tijms, H.C., 2003. A First Course in Stochastic Models. John Wiley and Sons.

Tovia, F., Brooks, R.M., Cassady, C.R., Rossetti, M.D., 2010. Modelling and analysis of service parts logistics systems. Int. J. Oper. Res. 10 (1), 60–81.

van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., De Boeck, L., 2013. Personnel scheduling: a literature review. Eur. J. Oper. Res. 226 (3), 367–385.

van Houtum, G.J., Kranenburg, B., 2015. Spare Parts Inventory Control under System Availability Constraints, vol. 227 Springer.

van Jaarsveld, W., Dollevoet, T., Dekker, R., 2015. Improving spare parts inventory control at a repair shop. Omega 57, 217–229.

van Jaarsveld, W., Scheller-Wolf, A., 2015. Optimization of industrial-scale assemble-to-order systems. Inf. J. Comput. 27 (3), 544–560.

van Wijk, A., Adan, I.J., van Houtum, G.J., 2012. Approximate evaluation of multi-location inventory models with lateral transshipments and hold back levels. Eur. J. Oper. Res. 218 (3), 624–635.

Vliegen, I.M.H., 2009. Integrated Planning for Service Tools and Spare Parts for Capital Goods. Ph.D. thesis. Technische Universiteit Eindhoven.

Vliegen, I.M.H., van Houtum, G.J., 2009. Approximate evaluation of order fill rates for an inventory system of service tools. Int. J. Prod. Econ. 118 (1), 339–351.

Waller, A., 1994. A queueing network model for field service support systems. Omega 22 (1), 35–40.

Wee, K.E., Dada, M., 2010. A make-to-stock manufacturing system with component commonality: a queuing approach. IIE Trans. 42 (6), 435–453.

Wong, H., van Houtum, G.J., Cattrysse, D., van Oudheusden, D., 2005. Simple, efficient heuristics for multi-item multi-location spare parts systems with lateral transshipments and waiting time constraints. J. Oper. Res. Soc. 1419–1430.

Zhao, Y., 2009. Analysis and evaluation of an assemble-to-order system with batch ordering policy and compound Poisson demand. Eur. J. Oper. Res. 198 (3), 800–809.

Zhao, Y., Simchi-Levi, D., 2006. Performance analysis and evaluation of assemble-to-order systems with stochastic sequential lead times. Oper. Res. 54 (4), 706–724.

Zhou, W., Chao, X., 2012. Stein–chen approximation and error bounds for order fill rates in assemble-to-order systems. Nav. Res. Logist. 59 (8), 643–655.