OR Spectrum (2017) 39:711–748
DOI 10.1007/s00291-017-0473-3

REGULAR ARTICLE

Integrated planning of spare parts and service engineers with partial backlogging

S. Rahimi-Ghahroodi[1] · A. Al Hanbali[1] · W. H. M. Zijm[1] · J. K. W. van Ommeren[2] · A. Sleptchenko[3]

Received: 29 February 2016 / Accepted: 10 January 2017 / Published online: 1 February 2017

Abstract In this paper, we consider the integrated planning of resources in a service maintenance logistics system in which spare parts supply and service engineers deployment are considered simultaneously. The objective is to determine close-to-optimal stock levels as well as the number of service engineers that minimize the total average costs under a maximum total average waiting time constraint. When a failure occurs, a spare part and a service engineer are requested for the repair call. In case of a stock-out at spare parts inventory, the repair call will be satisfied entirely via an emergency channel with a fast replenishment time but at a high cost. However, if the requested spare part is in stock, the backlogging policy is followed for engineers. We model the problem as a queueing network. An exact method and two approximations for the evaluation of a given policy are presented. We exploit evaluation methods in a greedy heuristic procedure to optimize this integrated planning. In a numerical study, we show that for problems with more than five types of spare parts it is preferable to use approximate evaluations as they become significantly faster than exact evaluation.

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.

✉ S. Rahimi-Ghahroodi
s.rahimighahroodi@utwente.nl

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

[2] Department of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

[3] Department of Mechanical and Industrial Engineering, College of Engineering, Qatar University, P.O. Box 2713, Doha, Qatar

Moreover, approximation errors decrease as problems get larger. Furthermore, we test how the greedy optimization heuristic performs compared to other discrete search algorithms in terms of total costs and computation times. Finally, in a rather large case study, we show that we may incur up to 27% cost savings when using the integrated planning as compared to a separated optimization.

Keywords Maintenance logistics · Queueing · Spare parts inventory · Field service · Approximate evaluation · Heuristic optimization

1 Introduction

Maintenance logistics is an important discipline that has received considerable attention both in practice and in the scientific literature. This attention is related to the often high investments associated with capital-intensive assets, which require a high operational availability. The unplanned downtime of advanced capital equipment can be extremely expensive. Consequently, these unplanned downtimes should be avoided as much as possible, and if they occur, they should be kept as short as possible (by using optimal corrective maintenance policies). The latter implies that malfunctioning parts or components causing the system breakdown are immediately replaced by ready-for-use ones since repair of a part on-site generally requires too much time. Such a policy in turn requires high availability of the resources (spare parts, tools, and service engineers) that are needed to execute corrective maintenance. However, these resources are mostly expensive and need high investments. This creates a large interest in cost savings; even savings of a few percent constitute a significant amount of money in absolute terms. Therefore, an optimal availability of resources in maintenance logistics is necessary to meet the expected operational availability while minimizing the total service costs. So far the planning of resources such as spare parts, service engineers, and repair tools has been mostly determined in isolation despite the fact that these resources have combined impacts on the performance of the system.

Unlike spare parts inventory management which is an indispensable element in maintenance logistics for any type of system, tools and service engineers are not always considered as bottlenecks in service logistics. In some cases, the replacement of failed parts can be done by operators in the production line. This makes the study of manpower availability unnecessary. Similarly, the required tools for the repair process are often cheap, and hence, every engineer has his own set of tools.

In this paper, we consider a service logistics system in which besides spare parts, highly skilled and trained service engineers are needed for the corrective maintenance. Since these highly skilled service engineers are considered as an expensive resource, manpower planning becomes necessary. In the current study, tools are no bottleneck for the system and considered to be always available. This paper focuses on the challenging multi-resource planning problem for spare parts and service engineers. The objective is to determine the required capacity of each resource minimizing the total service costs subject to a specified target on service level.

Before investigating the required capacity, it is important to discuss the competitive strategy of the service provider. Since both the spare parts and the number of service engineers are limited, we have to determine what policy should be followed if either resource is not immediately available. The two main strategies in any service policy are the cost-efficient strategy and the responsive strategy, which are in a sense two opposite extremes (see, e.g., Chopra and Meindl 2013). Depending on where a service provider wants to position its strategy between these two extremes, an appropriate service policy can be defined. On the one hand, when the objective is to have the highest service level (responsive strategy), waiting for resources is not acceptable. Then, a suitable policy is to use the emergency channel if any resource (either spare part or service engineer or both) is not available. On the other hand, when cost efficiency is the main objective, any additional cost (emergency shipment) should be avoided. In such a case, when a spare part or a service engineer is needed but there is no one available, the repair call has to wait until the needed resource becomes available via the conventional replenishment channel (backlogging policy).

In addition to these two extreme policies, various other policies can be applied. Usually, the nominal repair times of a system (i.e., excluding all extra waitings due to unavailable resources) are shorter than spare parts replenishment times. Therefore, in cases where a short waiting time is tolerable, queueing for service engineers is efficient if they are not immediately available. However, it is arguable to use the emergency shipment for spare parts in case of a stock-out. In this paper, we consider a full emergency policy (both for spare parts and for service engineers) in case of a spare part stock-out. If, however, the spare part is in stock but no service engineer is immediately available, a backlogging policy is followed for the latter.

The paper is organized as follows: First, we review the related literature in Sect. 2. In Sect. 3, we describe the model and different policies and scenarios are discussed. In Sect. 4, the model assumptions are introduced, and the model is investigated in detail. An exact and two approximation methods are proposed for the performance evaluations. To gain more insight into the model, numerical experiments are carried out to compare the results of the approximation methods with exact solutions. A heuristic to determine a near-optimal policy is developed in Sect. 5. We compare the optimization result of the proposed heuristic with other optimization algorithms. In Sect. 6, brief concluding remarks are summarized.

2 Literature review

2.1 Maintenance logistics

Maintenance logistics is a topic widely studied in the literature. One of the important areas in maintenance logistics that attracted a lot of attention is spare parts management. The amount of literature on (multi-item) spare parts optimization models is extensive and dates back to the pioneering paper of Sherbrooke (1968), who developed the METRIC (Multi-Echelon Technique for Recoverable Item Control) model. Sherbrooke (2004) and Muckstadt (2005) give a full overview of further developments from a methodology point of view in this area. Basten and van Houtum (2014) and van Houtum and Kranenburg (2015) discuss more recent models on spare parts inventory control with a focus on the system-oriented perspective.

In this research, we study the resource management of after-sales service logistics by considering both spare parts inventory management and manpower (service engineers) planning. Spare parts management and manpower planning in service logistics have been studied in isolation in many papers. The integration of spare parts and manpower planning was, however, rarely considered. The few papers that do study this integration mainly use simulation as a methodology for the performance analysis (Hertz et al. 2014; Visser and Howes 2007). Hertz et al. (2014) review the literature on simulation models in after-sales service logistics. Waller (1994) and Papadopoulos (1996) develop queueing network models for field service support systems considering manpower allocation and spare parts availability with emergency shipment options. Waller (1994) studies a model with only one service engineer. The service engineer carries a spare part kit, which can serve a fraction of all possible failures. If a spare part is not available in the kit, it is ordered from a depot and delivered directly to the customer. During this time, the service engineer visits other customers for repair. The availability of the part is known when the service engineer visits the customer. After the arrival of the part, the customer enters the waiting queue for the service engineers again. Customers are served by FCFS policy. The problem is modeled as a BCMP queuing network with two classes of customers, the ones waiting for an initial visit and the others waiting for a second visit after the arrival of the emergency delivered spare parts. Then, the model is used to evaluate different inventory and staffing policies. Papadopoulos (1996) extends this approach by considering multiple service engineers and introducing priority classes for customers via the application of the priority mean value analysis (PMVA) algorithm. He models the system as a closed queueing network.

Besides spare parts and service engineers, in a number of systems service tools are also needed to support the repair actions. Vliegen (2009) studies the integrated service tools and spare parts planning. In this research, the coupling of demand for tools and spare parts is considered explicitly. In addition, she also studies the coupling of tools in returns. This is a key difference with our model. She shows that integrating the planning of spare parts and repair tools leads to more accurate results and a cost savings of up to 15%.

Although there is a limited number of analytical models for the integrated spare parts management and manpower planning, there are quite a number of studies in other areas that can be used to help solving our problem. We review related literature in cross-trained manpower planning, assemble to order system, call center staffing and planning, and lateral transshipment inventory models.

2.2 Cross-trained manpower planning

In service logistics, one of the areas that has received considerable attention is the planning of skilled service representatives (the manpower) that are responsible for serving a number of service regions. In some papers, the field service system with dedicated and flexible (cross-trained) servers is studied. Usually, these papers consider the case that there are two or three different server types and one flexible team and use simulation to analyze the system (Agnihothri and Karmarkar 1992; Agnihothri et al. 2003; Agnihothri and Mishra 2004). Agnihothri and Karmarkar (1992) study the performance analysis of service territories by a queueing model and use simulation to test the accuracy. Agnihothri and Mishra (2004) examine service systems with cross-trained servers with two and three server types. The spare parts in our model can be seen as servers that have a specified skill and can serve one type of jobs only, and the service engineers can be seen as servers that can process any type of jobs (cross-trained servers). By this analogy, our problem is similar to the cross-trained manpower planning problem. In Brickner et al. (2010), a system similar to the one in this paper is modeled using simulation. They simulate a service system with three types of dedicated server teams and one flexible team. They assume a finite buffer for backorders and use a priority scheme to select the jobs from the buffer. They analyze the performance measurement of the system and then find the optimal number of each server type through a numerical search. In contrast to our model, there is no simultaneous request of servers in cross-trained manpower planning, so the model evaluation in these models differs. However, in the analytical papers, they use optimization approaches similar to ours to find the optimal number of servers.

2.3 Assemble to order system

Using different resources simultaneously for production orders (coupling in demand) is an aspect that makes our problem similar to assemble to order systems (ATO). In those systems, several subassemblies are demanded and all have to be available before an order can be processed. Song et al. (1999) study a generalized model that has both complete backlogging and lost sales as a special case. In addition, they distinguish total order service, which means that an order is fulfilled completely or rejected as a whole, and partial order service, which means that partial fulfillment is allowed. In Song et al. (1999), an exact performance analysis is carried out using matrix-geometric techniques that lead to a computationally efficient performance evaluation procedure. The supply system of each component is modeled as an independent production facility with a single exponential processor and a finite buffer, an M/M/1/c queue. Dayanik et al. (2003) study computationally efficient performance estimates for the same problem. Approximate models for base-stock assembly systems are also studied in Avsar et al. (2009). Hoen et al. (2011) develop an efficient and accurate approximation for an ATO system with deterministic lead times, where the lead times can be different for different items.

Song and Zipkin (2003) give an overview of papers on ATO systems. In most of the studies backlogging is assumed, but in some papers, the lost sales case is considered. In the ATO system literature, there are models where orders for various product types arrive stochastically to an ATO system. Each product type needs a set of components to be assembled. Lu et al. (2003) analyze such an ATO system as a set of queues driven by a common, multi-class batch Poisson input and derive the joint queue-length distribution. When comparing our model to these ATO models, we observe the same structure for demands, and the replenishment and service times in our model are like the replenishment lead times in an ATO system.

With regard to optimizing the stock levels in an ATO system, only a few papers consider lost sales. Benjaafar and ElHafsi (2006) study the optimal policy for the base-stock levels of components used in a single end-product. ElHafsi et al. (2008)

extend the model of Benjaafar and ElHafsi (2006) to a situation with multiple products. However, their analysis is restricted to a nested design, i.e., product i has only one additional component more than product $i-1$.

2.4 Call center staffing and planning

By treating the spare parts as servers (in addition to service engineers) and considering the spare parts replenishment time as a service time, there are quite a number of papers in the call center area that study similar models. Mostly, they use queueing modeling. Since call center systems in practice face high traffic and the number of servers is high, an asymptotic analysis of the call center is often performed. Usually, in the after-sales service logistics the number of service engineers is not that high, so the asymptotic results are not useful.

Generally, for problems where there are multi-type customers in call center systems and servers have different skills, similar approaches can be observed as we use in this paper. However, as for cross-trained manpower planning, there is no simultaneous demand for servers in call center models. A survey paper in this area is done by Koole and Pot (2006) who review the staffing and routing problem of multi-type customers in a call center. Shumsky (2004) studies an approximation model for a service system with two dedicated servers and one flexible server by using a queuing model. He provides an estimation for performance measurement of a call center system. Ormeci (2004) models a Markovian loss system for a call center with two different customer classes with different revenue and service and arrival rates. There are three different servers, two dedicated for each customer type and one flexible server that can serve both customer types, where the dedicated servers work faster than the flexible one. She shows that serving a call in its dedicated station, whenever possible, is optimal. For the shared station, since the customers have different priorities (revenue), there exists an optimal monotone threshold policy.

Spare parts can be named as dedicated resources since for each repair call a specific spare part is needed while the service engineers are shared for all types of repair calls. There are a limited number of studies in which a service system with combination of shared and dedicated resources is analyzed. Akşin and Harker (2003) consider an inbound call center system with multi-type customers served by dedicated servers and one shared resource (IT infrastructure). The shared resource in the system is treated as a process sharing server. Due to the specific call center system operations, there is a fundamental difference between our model and that of Akşin and Harker (2003). Namely, for each call a dedicated server and shared resource are needed but by finishing the call, both server and shared resource will be free simultaneously.

2.5 Lateral transshipment inventory models

The approximate evaluation methods that we provide are related to the approximate evaluation methods that are proposed in lateral transshipment inventory models. In these systems, to determine the optimal policy, evaluation of costs of a given setting is necessary. For this, the stream of lateral transshipment requests between the warehouses is commonly approximated by Poisson processes (Axsäter 1990; Alfredsson and Verrijdt 1999; Kukreja et al. 2001; Kutanoglu 2008; Van Kranenburg and Houtum 2009). However, approximating overflow processes in lateral transshipment models (or similarly accepted arrival processes in our model) with Poisson processes is not always reliable. van Wijk et al. (2012) perform an extensive numerical study and show that Poisson approximations do not always give satisfactory accuracy. Here, we propose other fast approximation methods that give more accurate result than using Poisson arrival processes. van Wijk et al. (2012) propose a new approximation algorithm for the evaluation of a given policy, using interrupted Poisson processes (IPP, cf. Kuczura 1973) that is more accurate but computationally more expensive.

Greedy algorithms are commonly used for optimization in lateral transshipment inventory models. Wong et al. (2005) propose a greedy method with a local search for multi-item multi-location spare parts systems with lateral transshipments and waiting time constraints. Van Kranenburg and Houtum (2009) exploit a similar greedy algorithm without any local search for the optimization of their partial pooling structure in spare parts networks. In both papers, the authors show that the greedy algorithm performs reasonably well.

Overall, the literature study indicates that the multi-resource planning in maintenance service logistics so far lacks a thorough analysis. In the next section, we develop a model for the integrated spare parts and service engineers planning. We take advantage of existing models in other applications which we discussed above. In particular, the evaluation procedures and queueing models that are used in ATO, call center, and lateral transshipment models help us to develop our evaluation methods. Moreover, in our optimization problem and algorithm, we use some concepts and techniques that are presented in the spare parts inventory management and cross-trained manpower planning literature.

3 Model description

We consider a service region with a local inventory to store K different types of spare parts. There are different types of repair calls in this service region that arrive randomly with a rate λ. Each repair call requires a specific spare part. A repair call is of type-k if it requires one unit of type-k spare part. Let p_k denote the probability that a repair call is of type k. For each repair call, a service engineer is also needed to do the job. A team of service engineers is located in the service region. In this model, we assume that the service time of a repair call of type k (the time between the moment the repair job is assigned to a service engineer and the moment the job will be finished) is exponentially distributed with rate μ_k. The inventory of type-k parts is managed according to the base-stock policy, referred to as $(S_k - 1, S_k)$. For each type-k spare part, the replenishment lead time is exponentially distributed with rate ν_k. For each spare part, there is a holding cost per item per time unit. In addition, hiring costs of service engineers and emergency costs (a cost per repair call that is satisfied via an emergency channel) are considered in this model.

Depending on the importance of service level and the height of the downtime cost, different service policies can be followed. Consider a system that, if a failure happens, should be fixed as soon as possible while the downtime cost is much higher than the holding and transshipment costs. In this case, the service policy should apply an emergency channel for both spare parts and service engineers in case there is a shortage of any of these resources. So, upon a request arrival, if any of the needed spare part or service engineer is not available, the repair call is considered to be lost for the internal system and both the spare part and the service engineer are satisfied via an external emergency channel with a high cost. However, for systems for which downtime does not cost much or failures will not stop the whole system, we can assume backlogging for both spare parts and service engineers. So, when one of them is missing, we just wait until a spare part or a service engineer becomes available.

The aforementioned scenarios are two extreme strategies. In practice, there are a variety of policies in between that can be applied. Here, we are not interested in the full emergency (the most responsive strategy) nor in the full backlogging policy (the most cost-efficient strategy). Instead, we study a more cost-efficient strategy that has less effect on the waiting times. Usually, service times take much less time than the spare parts replenishment. Therefore, the first step to make the most responsive strategy more cost-efficient is by changing the service engineers policy to backlogging. In other words, in systems for which a short waiting time is acceptable, it is rational to wait for service engineers if they are not immediately available but use the emergency shipment for spare parts in case of a stock-out. That is the scenario we study in this paper.

When the requested spare part is satisfied by an emergency shipment, there are a number of scenarios for the service engineers that can be applied. In this paper, we assume that both spare parts and service engineers are satisfied via an emergency channel in case of a spare parts stock-out. In other words, internal service engineers are not responsible for emergency repair calls. To explain why this assumption is justified, let us discuss the possible scenarios where just spare parts are satisfied by emergency channel. First, suppose the repair request is sent to service engineers after receiving the spare part emergency shipment. In this case, for systems where service engineers traveling time to the failure location is a considerable amount of the total service time (it usually includes service engineers travel time and on-site repair time), this scenario causes extra waiting time for the system and the service engineers always arrive later than spare parts emergency shipments. On the contrary, if the pool of service engineers receives the request already when the failure happens (in the case where the spare part is going to be satisfied by an emergency channel), the service engineer may arrive to the location sooner than the emergency shipment, which causes extra waiting time and decreases the service engineer's utilization. Furthermore, in practice, there are cases where the spare parts are transferred by the service engineers to the failure locations. It results in lower shipment costs and ensures that service engineers and the spare parts will arrive at the same time. By outsourcing the repair job to the external service engineers (when the spare part is delivered via the emergency channel), the spare part and the service engineer will arrive in the failure location at the same time. More precisely, suppose the emergency shipments are sent from a central station (depot) where service engineers are always available. So, when the spare part is satisfied by an emergency shipment, a central service engineer goes directly together with the requested spare part to the failure location to perform the repair job. Hence, when the requested spare part is not available in stock, the repair call is satisfied entirely via the emergency channel. After all, the other scenario where only spare parts are requested by emergency shipments is also justified in practice and will be studied in future work. However, it is less responsive but at the same time, we expect less total cost.

When the spare part is available, the backlogging policy is followed for service engineers. When no service engineer is available upon a repair call request (while the spare part is available), the spare part is reserved and the system must wait until a service engineer becomes available. A maximum accepted average waiting time is defined for the total waiting time in the service region (not per spare part), and there is no priority over different spare part types. Therefore, the backorders in the service engineers queue will be served by a FCFS policy.

All in all, we are interested to find the optimal spare parts stock levels and the optimal number of service engineers to minimize the total average costs (holding, service engineers hiring, and emergency costs) under a maximum total average waiting time constraint. Waiting times are caused by the emergency shipments and by queueing for service engineers. First, let us summarize some notations.

Notation:

Spare parts:	$k : 1, \ldots, K$;
λ:	Total failure rate in the service region;
p_k:	Probability that the repair call needs type-k spare part;
ν_k:	The regular replenishment rate for type-k spare part;
ν_k^{em}:	The emergency replenishment rate for type-k spare part;
μ_k:	Service rate for type-k repair job (i.e., the reciprocal of the repair time);
S_k:	Stock level for type-k spare part;
E:	Number of service engineers;
W^S:	Expected waiting time of calls in the spare parts inventory (waiting for spare parts occurs in case of an emergency shipment);
W^E:	Expected waiting time of calls in the service engineers queue;
W:	Total expected waiting time of all repair calls in the service region.

In the next section, we model the problem based upon the aforementioned scenario where we have a (full) emergency replenishment in case of a depleted local spare parts stock and backlogging for service engineers.

4 Performance evaluation

In this section, we describe how a given policy, i.e., a choice of all base-stock levels and the number of service engineers, can be evaluated, either exactly or approximately. The service policy is defined as follows. In the case that a repair call arrives in the service region and the requested spare part is not available, the repair call will be satisfied entirely (both the needed spare part and the service engineer) by an emergency channel with a high cost. However, when there is no available service engineer while the spare part is available, this spare part will be reserved and the repair call will be backlogged until a service engineer becomes available. Backorders are served according to FCFS policy.

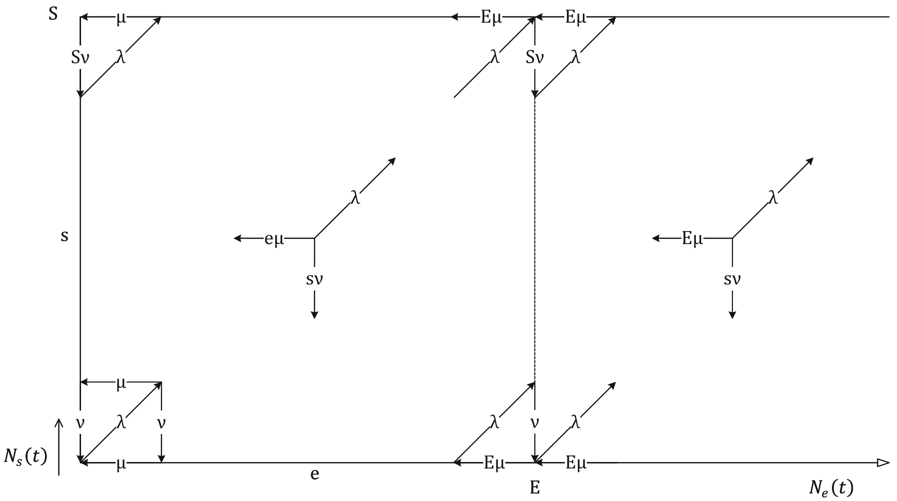

Fig. 1 Transition diagram of the model with full spare parts emergency channel and repair backlogging when there is one type of spare part ($K = 1$)

4.1 Exact evaluation with Markov chain

Let us denote $N_{s_k}(t)$ as the number of type-k spare parts in the pipeline (replenishment), and $N_e(t)$ as the number of calls waiting or being served in the service engineers queue. Under the above assumptions, the process $N(t) = \big(N_e(t), N_{s_1}(t), N_{s_2}(t), \dots, N_{s_K}(t); t \geq 0\big)$ is a continuous-time Markov chain with the following infinite size state space

$$\Omega = \{0, \dots, E, \dots, \infty\} \times \{0, \dots, S_1\} \times \{0, \dots, S_2\} \times \cdots \times \{0, \dots, S_K\}. \quad (1)$$

This Markov chain can be analyzed using the matrix-geometric method, but for large K, the numerical evaluation will be computationally expensive, if not intractable. More precisely, the dimension of the rate matrix in the matrix-geometric method increases exponentially with $S_k, k = 1, \dots, K$, and with K. For small size problems, we explain how to find the steady-state probabilities using the matrix-geometric method. In Fig. 1, we show the transition rate diagram for the Markov chain in the simple case with one type of spare part ($K = 1$) with stock level S, and a team of engineers with size E.

In this part, the matrix-geometric method for this problem is explained. For the sake of simplicity, we assume that the service rate is the same for all types of spare parts. Using the matrix-geometric method is also possible in case of non-equal service rates. However, the formulation will be more complex. For the joint process $N(t)$, we shall refer to $N_e(t)$ as the level of the process and $\big(N_{s_1}(t), N_{s_2}(t), \dots, N_{s_K}(t)\big)$ as the phase. Let us introduce the following matrices. Let $U_k, k = 1, \dots, K$, denote an upper diagonal $(S_k + 1, S_k + 1)$ matrix with upper diagonal elements equal to $\lambda_k = p_k\lambda$. Let $B_k, k = 1, \dots, K$, denote a bi-diagonal lower $(S_k + 1, S_k + 1)$ matrix with j-th main diagonal element $-(j-1)\nu_k - \lambda_k 1_{\{j \leq S_k\}}$ for $j = 1, \ldots, S_k + 1$, where $1_{\{j \leq S_k\}}$ is the indicator function, and lower diagonal j-th element $j\nu_k$ for $j = 1, \ldots, S_k$.

The process $N(t)$ is a level-dependent quasi-birth-death process for levels $0, \ldots, E$, and level independent for the rest with the generator G given by:

$$G = \begin{bmatrix} A_1^0 & A_2 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ A_0^1 & A_1^1 & A_2 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & \ddots & 0 & \ddots & \ddots & \ddots & \ddots \\ \vdots & 0 & A_0^{E-1} & A_1^{E-1} & A_2 & 0 & \ddots & \ddots & \ddots \\ \vdots & \ddots & 0 & A_0^E & A_1^E & A_2 & 0 & \ddots & \ddots \\ \vdots & \ddots & \ddots & 0 & A_0^E & A_1^E & A_2 & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}, \tag{2}$$

where $A_0^l = \mu l I$, $A_1^l = B_1 \oplus B_2 \oplus \cdots \oplus B_K - \mu l I$, $l = 0, \ldots, E$, and $A_2 = U_1 \oplus U_2 \oplus \cdots \oplus U_K$. $A \oplus B$ is the Kronecker sum of A and B and is equal to $A \otimes I + I \otimes B$. $A \otimes I$ is the Kronecker product of A and identity matrix, I.

Let $V = (V_0, V_1, \ldots)$ denote the steady-state probability of $N(t)$, i.e., $VG = 0$ with $Ve^T = 1$; e^T is a column vector of entries equal to one. The balance equations are as follows:

$$V_0 A_1^0 + V_1 A_0^1 = 0, \tag{3}$$

$$V_{i-1} A_2 + V_i A_1^i + V_{i+1} A_0^{i+1} = 0, \quad i = 1, \ldots, E-1, \tag{4}$$

$$V_{i-1} A_2 + V_i A_1^E + V_{i+1} A_0^E = 0, \quad i = E, E+1, \ldots. \tag{5}$$

The general solution of V_i is of type $V_{i-1} R_{\min(i,E)}$, $i = 1, 2, \ldots$ (for details, see Neuts 1981). Then, Eqs. (4) and (5) give

$$R_E = -A_2 \left(A_1^E + R_E A_0^E \right)^{-1}, \tag{6}$$

$$R_l = -A_2 \left(A_1^l + R_{l+1} A_0^{l+1} \right)^{-1}, \ l = E-1, \ldots, 1. \tag{7}$$

Equation (6) can be solved using a standard iterative procedure of the matrix-geometric methods (see, e.g., Latouche and Ramaswami 1999). When R_E is known, all R_l can be computed using the backward iteration defined in (7).

Inserting the general solution of V_1 in Eq. (3) yields $V_0(A_1^0 + R_1 A_0^1) = 0$. The latter equation together with the normalization condition

$$\sum_{i=0}^{\infty} V_i e^T = V_0 \left(I + R_1 + R_1 R_2 + \cdots + R_1 R_2 \ldots R_{E-1} (I - R_E)^{-1} \right) e^T = 1 \tag{8}$$

gives V_0, then $V_1 = V_0 R_1$, $V_2 = V_1 R_2$, and so forth.

Based on the steady-state probabilities, the required performance measures can be determined. For example, the expected number of repair jobs waiting in the service engineers queue is given by

$$\begin{aligned} Q^E &= \sum_{i=1}^{\infty} i \ V_{E+i} \ \ e^T \\ &= \sum_{i=1}^{\infty} i \ V_E (R_E)^i \ \ e^T \\ &= V_E \sum_{i=1}^{\infty} i \ (R_E)^i \ \ e^T \\ &= V_E R_E (I - R_E)^{-2} \, e^T. \end{aligned} \tag{9}$$

Given the expected number of repair jobs waiting in the service engineers queue, we can derive the expected waiting times using Little's law.

Note that the inversion of matrices in the computation of R_i, $i = 1, \ldots, E$, is of complexity $\prod_{k=1}^{K} (S_k + 1)^3$. So, we conclude that the total complexity to find the steady-state probabilities is equal to

$$(E + 1) \prod_{k=1}^{K} (S_k + 1)^3. \tag{10}$$

4.2 Approximate evaluation

Performing an exact evaluation for this problem will not be efficient for large size problem. Hence, we develop a more efficient method to obtain an approximate result. In this method, the model evaluation is done in two steps. In Sect. 4.2.1, we examine the spare parts inventory to find the emergency rate and average waiting time that is caused by emergency shipment. Then, given the spare parts inventory evaluation, we analyze the service engineers queue in Sects. 4.2.2 and 4.2.4. Note that the results in Sect. 4.2.1 are exact while we use approximation methods for service engineers queue.

4.2.1 Emergency rate and average waiting time in spare parts inventory

As mentioned before, we have an emergency shipment (loss) system for spare parts and a backlogging system for service engineers, as long as spare parts are available. We assign the spare part when a repair call arrives, whether a service engineer is available or not. The emergency probability for repair calls is defined as the fraction of calls that will be satisfied by the emergency shipment. Note that this probability is only a function of spare parts stock levels.

In the following, we show how the number of type-k spare parts in the replenishment pipeline can be modeled as the number of jobs in an $M/M/S_k/S_k$ queue. For any spare part of type k, there are S_k spare parts, that can be seen as servers. The repair calls arrive according to a Poisson process with rate $\lambda_k = p_k\lambda$. When a spare part has been allocated to a call (becomes busy), it takes an exponential time to replenish it by a new part (service time) with rate ν_k. When there is no spare part in the inventory (all servers are busy), the arriving calls will be lost and satisfied by the emergency channel. It means the maximum number of parts in replenishment in this queueing model is equal to S_k. Let $\rho_k = \lambda_k/\nu_k$. The emergency probability in an $M/M/S_k/S_k$ (using PASTA property) is given by Erlang B (loss) formula;

$$P_k^L(S_k) = \frac{\rho_k^{S_k}/S_k!}{\sum_{i=0}^{S_k} \rho_k^i/i!}. \tag{11}$$

The emergency rate of type-k repair calls as a function of type-k spare parts stock level is equal to

$$\lambda_k^L(S_k) = \lambda_k P_k^L(S_k). \tag{12}$$

In this model, we assume that the emergency replenishment rate is much higher than the regular one, but still finite. It means when a repair call is satisfied by an emergency channel, there is still a waiting time until the emergency shipment arrives. This waiting time is important for the service policy and is included in the maximum accepted average waiting time of repair calls. For the parts that are requested by emergency shipment, the average waiting time is equal to $1/\nu_k^{\text{em}}$. Note that the average (emergency shipment) waiting time of all repair calls is equal to a weighted sum of the repair calls that are satisfied by emergency shipment times $1/\nu_k^{\text{em}}$, as follows

$$W^S(\mathrm{S}) = \sum_{k=1}^{K} \frac{p_k P_k^L(S_k)}{\nu_k^{\text{em}}}. \tag{13}$$

where $\mathrm{S} = \{S_1, \ldots, S_K\}$ is the vector of base-stock levels.

4.2.2 Average waiting time in service engineers queue—MVA approximation

In this section, we are interested in finding the average waiting time of repair calls that is caused by the limited number of service engineers. When there is no available spare part, the call is entirely served externally and hence is lost for the internal system. Therefore, the arrival rate as experienced by the service engineers queue equals

$$\gamma = \sum_{k=1}^{K} \gamma_k = \sum_{k=1}^{K} \lambda_k \left(1 - P_k^L(S_k)\right). \tag{14}$$

Note that arrival streams to the service engineers queue (each arrival stream is related to one type of repair call) are not renewal processes. Upon arrival, when there is on-hand spare parts inventory, the call will be forwarded immediately to the engineers queue. However, when the spare parts inventory become empty, calls are satisfied by the emergency channel. This dependency of arrivals on spare parts stock inventory causes arrivals at the internal service engineers pool to be dependent on past arrivals and makes arrival streams non-renewal processes. The correlation between inter-arrival times depends on the spare parts base-stock level. When the base-stock level is zero or very large, we can say that there is no correlation between inter-arrival times. For small values of the base-stock level, there is a correlation between inter-arrival times. However, we have tested the correlation numerically, and we found it almost negligible. The inter-arrival correlation in different situations was always below 0.05. Nevertheless, the total arrival process is still a non-renewal process. Apart from that, there is no correlation between different types of repair call arrivals to the service engineers queue. The arrival streams to the spare parts inventory are independent Poisson processes. Moreover, since in each repair call, just one specific spare part is required, there is no dependency between stocks of different spare parts types. Hence, the arrival streams for different types of repair calls to the service engineers queue are independent. We show in the numerical result section that this independency causes our approximation method to become more accurate when there are more arrival streams (more spare part types).

We have a multi-class multi-server queue for service engineers with total arrival rate γ and service rate μ_k for the repair call of type k, $k = 1, \ldots, K$. So, the service times in the service engineers queue follow a hyper-exponential distribution with rate η which is given by

$$\frac{1}{\eta} = \sum_{k=1}^{K} \frac{\alpha_k}{\mu_k}, \tag{15}$$

where α_k is the probability that the repair call that has arrived to the service engineers queue is of type k, which gives

$$\alpha_k = \frac{\gamma_k}{\gamma}. \tag{16}$$

We have a $G/H/E$ queue for the service engineers queue (H refers to the hyper-exponential service time). In the $G/H/E$ queuing system, no exact results are available for the mean waiting time, but the mean value analysis (MVA) approach can be used heuristically to derive a simple approximation (see Tijms 2003). First, we need to have the probability that all servers are busy. Define $\sigma = \frac{\gamma}{\eta}$ as the offered load. As an approximation, we can use the busy probability of an $M/M/E$ queue (see, e.g., Tijms 2003).

$$P^B = \frac{\frac{\sigma^E}{E!}}{(1 - \frac{\sigma}{E}) \sum_{j=0}^{E-1} \frac{\sigma^j}{j!} + \frac{\sigma^E}{E!}}. \tag{17}$$

The following formula gives us the average waiting time (excluding service time) for an $M/M/E$ queue:

$$\frac{P^B}{\eta(E - \sigma)}. \tag{18}$$

We do not expect that the average waiting time of an $M/M/E$ queue is a reliable approximation for the average waiting time in the service engineers queue. Therefore, to have a better approximation, we consider the coefficient of variations of the inter-arrival time and the service time in our formulation. Note that the arrival processes to the service engineers queue do not form a renewal process. However, from now on, we assume it is a renewal process in our approximation method. For the cases that the service rate per call type-k (μ_k) is not the same for different spare part types, the service time is also not exponentially distributed. By knowing the coefficient of variations of the inter-arrival time and the service time, a better approximation for the average waiting time is:

$$W^E_{\text{mva}}(\text{S}, E) = \left(\frac{c_s^2 + c_a^2}{2}\right) \frac{P^B}{\eta(E-\sigma)}. \tag{19}$$

where c_s^2 and c_a^2 are the squared coefficient of variation of the service time and the inter-arrival time, respectively. Note that the average waiting time in the service engineers queue is a function of all base-stock levels, S, and the number of service engineers, E.

4.2.3 Coefficient of variations

The service times in the service engineers queue follow the hyper-exponential distribution. Therefore, its coefficient of variation is equal to

$$c_s^2 = \frac{2\sum_{k=1}^{K} {}^{\alpha_k}/_{\mu_k^2}}{\left(\sum_{k=1}^{K} {}^{\alpha_k}/_{\mu_k}\right)^2} - 1. \tag{20}$$

To find the coefficient of variation of the arrival process, we need to analyze the type-k arrival process to the service engineers queue. Let us denote X_k as the inter-arrival time of type-k parts in the service engineers queue. We can show that X_k has a phase-type distribution with the following Laplace–Stieltjes transform function (for proof, see "Appendix 1").

$$\widetilde{X_k}(\omega) = \frac{\lambda_k\left(S_k\nu_k + (1-d_k)\omega\right)}{(\lambda_k+\omega)(S_k\nu_k+\omega)}, \tag{21}$$

where d_k represents the probability that no spare part of type-k is left in the inventory given that a type-k repair call has just been accepted. So, for the next arrival, first a replenishment must happen and then the next arrival will be accepted. Let $\pi_k(i)$, $i = 0, \ldots, S_k$ denote the steady-state probabilities in the type-k spare parts queue, i.e., the steady-state probabilities in an $M/M/S_k/S_k$ queue with arrival rate λ_k and service rate ν_k. The probability d_k is given by

$$d_k = \frac{\pi_k(S_k-1)}{1-\pi_k(S_k)} = \frac{\nu_k S_k P_k^L(S_k)}{\gamma_k}. \tag{22}$$

Now, we are able to find the type-k parts arrival process mean and variance using Laplace–Stieltjes transform function in (21) which are given by

$$E(X_k) = \frac{\lambda_k d_k + S_k\nu_k}{\lambda_k S_k \nu_k} = \frac{1}{\gamma_k}, \tag{23}$$

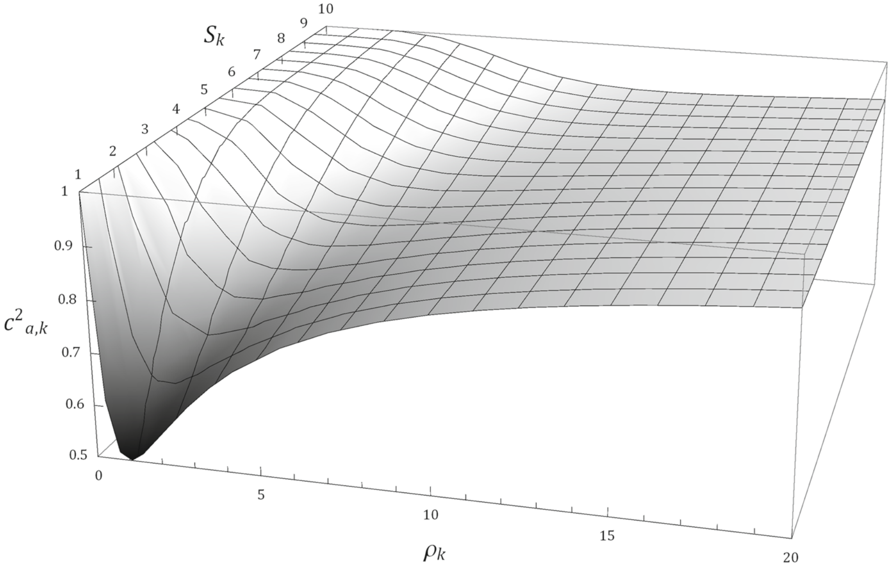

Fig. 2 Squared coefficient of variation of a single arrival stream versus ρ_k and S_k

$$\mathrm{Var}(X_k) = \frac{1}{\lambda_k^2} + \frac{d_k(2-d_k)}{S_k^2 \nu_k^2}. \tag{24}$$

So, the squared coefficient of variation for the type-k arrival process to the service engineers queue is equal to

$$c_{a,k}^2 = \gamma_k^2 \mathrm{Var}(X_k) = 1 - 2P_k^L + \frac{2\rho_k}{S_k}(1 - P_k^L)P_k^L. \tag{25}$$

Note that P_k^L is a function of ρ_k and S_k (see Eq. 11). Therefore, the squared coefficient of variation for the type-k arrival process is a function of only ρ_k and S_k. Figure 2 shows how $c_{a,k}^2$ changes as a function of ρ_k and S_k. As can be seen in the figure, $c_{a,k}^2$ is always between 0.5 and 1. It reaches its minimum when both ρ_k and S_k are equal to 1. When either ρ_k or S_k goes to infinity, $c_{a,k}^2$ converges to 1. By contradiction, it is easy to show that $c_{a,k}^2 < 1$.

In the literature, several approximation methods for the coefficient of variations of a superposition of arrival processes are introduced (see, e.g., Albin 1984; Whitt 1983). None of these methods gives an accurate result for the performance evaluation in our problem. Some of the approximations in the literature work good when the arrival processes are renewal (we have non-renewal arrival processes) or when the coefficient of variation of arrival processes is larger than one. Here, we design the following method to get an approximate coefficient of variation of the superposition process in this problem. In this method, we approximate the superposition stream as a superposition of identical streams with a Coxian-2 inter-arrival times. In the literature, an exact method is proposed to find the coefficient of variation of the superposition process of identical Coxian-2 arrival streams (see, e.g., van Vuuren 2007, p. 23). First let us denote

$$L_j = \sum_{k=1}^{j} \alpha_k c_{a,k}^2. \tag{26}$$

Now, first suppose we have two types of spare parts. Then, the arrival process to the service engineers queue is a superposition of two arrival streams. By approximating these two arrival streams with two identical Coxian-2 arrival streams, we obtain the equation below as an approximation for the coefficient of variation of the superposition of the two arrival processes (for a proof, see "Appendix 3").

$$c_a^2 = \frac{L_2(2 + L_2)}{1 + 2L_2}. \tag{27}$$

For 3 part types, in a similar way, we find that

$$c_a^2 = \frac{L_3(3 + 6L_3 + L_3^2)}{1 + 5L_3 + 4L_3^2}. \tag{28}$$

For problems with more than 3 types of spare parts, we can use the approach as explained in "Appendix 3" or in more details in van Vuuren (2007). However, as an alternative, we propose the following computational efficient iterative procedure. First, we replace all arrival streams with the same number of identical Coxian-2 arrival streams with the coefficient of variation given in (26). Then, for each two or three streams of arrival processes, we use Eqs. (27) or (28) and replace them with one arrival stream. We repeat this procedure until we find the coefficient of variation of the total arrival process.

Now, we have an approximation for the coefficient of variation of the arrival process to the service engineers queue. Therefore, we can use Eq. (19) as an approximation for the average waiting time in the service engineers queue.

4.2.4 Average waiting time in service engineers queue—LT approximation

In the previous section, we have shown that the arrival process for the service engineers queue is a superposition of independent phase-type (non-renewal) processes. Here, by using the Laplace transform (LT) of the arrival process, we propose another approximation method for the average waiting time based on the exact solution for the $GI/M/E$ queue (see, Takács 1962). We will show numerically that the MVA approximation method works poorly when there is a small number of spare part types or the emergency probability is rather high. In these cases, we can use the LT method that we introduce in this section. First we explain this method for problems with only one type of spare part in which there is no superposition of arrival processes.

Suppose ω^* is the root of the equation below in region (0, 1).

$$\widetilde{X}\big(E\eta(1-\omega)\big) = \omega, \tag{29}$$

where $\widetilde{X}(\omega)$ is given in (21). The solution of the previous equation is given by

$$\omega^* = \frac{\lambda + E\eta + S\nu - \sqrt{(E\eta + S\nu - \lambda)^2 + 4E\eta\lambda(1 - d_1)}}{2E\eta}. \tag{30}$$

Equation (31) gives the exact average waiting time in $GI/M/E$ queues where the arrival process is renewal.

$$W_{GI/M/E} = \frac{D}{E\eta(1-\omega^*)^2}, \tag{31}$$

where

$$D = \left[\frac{1}{1-\omega^*} + \sum_{j=1}^{E} \frac{\binom{E}{j}}{C_j(1-\widetilde{X}(j\eta))}\left(\frac{E(1-\widetilde{X}(j\eta))-j}{E(1-\omega^*)-j}\right)\right]^{-1}, \tag{32}$$

and

$$C_j = \prod_{i=1}^{j} \frac{\widetilde{X}(j\eta)}{1-\widetilde{X}(j\eta)}, \qquad j = 1, \ldots, E \tag{33}$$

Since the arrival process to the service engineers queue is not renewal and the service time is not exponentially distributed (when the service rate is not the same for different spare parts), Eq. (31) does not give the exact solution for the average waiting time even when there are one type of spare part and one service engineer. The only case in which the expression is exact is when we have one part type with the stock level equal to one. In this case, we know that the arrival process is renewal and the service time is exponential, so this method gives us the exact average waiting time. However, we use this method as an approximation for cases with non-renewal arrival process, where there are more than one spare part type, and the service time distribution is hyper-exponential. To make this approximation more accurate for the cases where the service time is hyper-exponentially distributed, we scale the average waiting time based on the coefficient of variation of the service time. Therefore, the equation below gives a simple approximation for the average waiting time in the service engineers queue.

$$W_{lt}^{E}(\mathrm{S}, E) = \left(\frac{1+c_s^2}{2}\right)\frac{D}{E\eta(1-\omega^*)^2}, \tag{34}$$

where c_s^2 is the squared coefficient of variation of the service time which is given in (20).

When there is more than one type of spare part, we need to find the Laplace transform of the superposition of the various arrival processes. Although finding the exact Laplace transform of the total arrival process is possible, we end up with a complex formulation. Moreover, to find the root of Eq. (29), we need a numerical search that may be computationally expensive when the number of spare part types increases. Therefore, for multi-part problems, we use the first two moments of the total arrival process to fit a simple distribution. To find a suitable option to which we fit the total arrival process, we need to know in what range the coefficient of variation of the inter-arrival times belongs. We observe that the coefficient of variation of the total arrival process is always between 0.5 and 1 (similar to the individual arrival streams). Therefore, we choose a Coxian-2 distribution for fitting the superposition of the arrival processes.

Suppose γ is the rate of the total arrival process to the service engineers queue and c_a is the coefficient of variation of that process. Equation (35) gives the Laplace transform of the fitted Coxian-2 distribution for the total arrival process, see "Appendix 2" for the proof,

$$\widetilde{X}(\omega) = \frac{\gamma\left(2\gamma + (2c_a^2 - 1)\omega\right)}{(\omega + 2\gamma)(\omega c_a^2 + \gamma)}. \tag{35}$$

The root ω_{co}^* of the equation $\widetilde{X}\left(E\eta(1-\omega)\right) = \omega$ (in (0, 1)) is given by

$$\omega_{co}^* = \frac{\gamma + 2c_a^2\gamma + E\eta c_a^2 - \sqrt{\left(\gamma - 2c_a^2\gamma + c_a^2 E\eta\right)^2 + 4c_a^2\gamma E\eta}}{2c_a^2 E\eta}. \tag{36}$$

Similar to the single part problem, Eq. (34) gives an approximate value for the average waiting time in the service engineers queue with now ω^* replaced by ω_{co}^*.

In the numerical section, we compare this approximation method with the MVA approximation for different parameters settings. One may think of other types of approximation like the two-moments approximation of Tijms (2003) for which performance measures of the $GI/D/C$ queue are needed (which are not known in closed form).

In summary, we propose two approximation methods to calculate the average waiting time in the service engineers queue. Note that W^E (W_{mva}^E or W_{lt}^E) gives us the expected waiting time of calls that arrive at the service engineers queue. To have the expected waiting time of all repair calls related to the service engineers queue, we should multiply it by the fraction of calls that are sent to the service engineers queue. So, the total expected waiting time in the system as a function of spare parts stock levels and the number of service engineers is equal to

$$W(\mathrm{S}, E) = \frac{\gamma}{\lambda} W^E(\mathrm{S}, E) + W^S(\mathrm{S}), \tag{37}$$

where W^S, defined in (13), is the average waiting time in the spare parts inventory that is caused by the emergency shipment, and W^E is the average waiting time in the service engineers queue that is defined approximately in (19) or (34) using the MVA and LT evaluation methods, respectively. It can be also determined exactly by dividing the expression in Eq. (9) by γ (14). Note, S is the vector of all spare parts stock levels.

4.3 Numerical comparison

In this subsection, we validate our approximate evaluation methods. Note that the average waiting time in the spare parts inventory (W^S) that we obtain in Sect. 4.2.1, Eq. (13) is exact. So, to validate our approximation method, we compare the approximate and the exact solutions of the average waiting time in the service engineers queue. First, we test the MVA approximation method in instances with five types of spare parts. We consider different parameter settings and examine the instances where the total emergency probability is less than 10%. We generate a number of instances randomly where the service rates are the same for all part types, $\mu_k = \mu, \forall k$, and vary from 1.5 to 9.6 calls per week. The total arrival rate, λ, is equal to 5 calls per week. The stock level, S_k, varies from 1 to 5 units, and the number of engineers, E, changes from 1 to 5. The replenishment rate increases from 1.2 to 9.6 parts per week.

Figure 3 shows the accuracy of the MVA approximation method (maximum and average approximation error) as a function of the number of service engineers and the total emergency probability. The approximation error is shown as a percentage in the figure and is calculated as $100 \times \frac{W^E_{\text{mva}} - W^E}{W^E}$, where W^E_{mva} is the approximate average waiting time (MVA) as given in (19) and W^E is the exact average waiting time derived from Eq. (9). In all instances, the approximation error is positive ($W^E_{\text{mva}} > W^E$). As can be seen in the figure, the approximation error increases with the total emergency probability. We know that the approximation method performs better when the arrival process to the service engineers queue is renewal or, in a loose sense, closer to a renewal process. So, the smaller the emergency probability, the better the approximate result is.

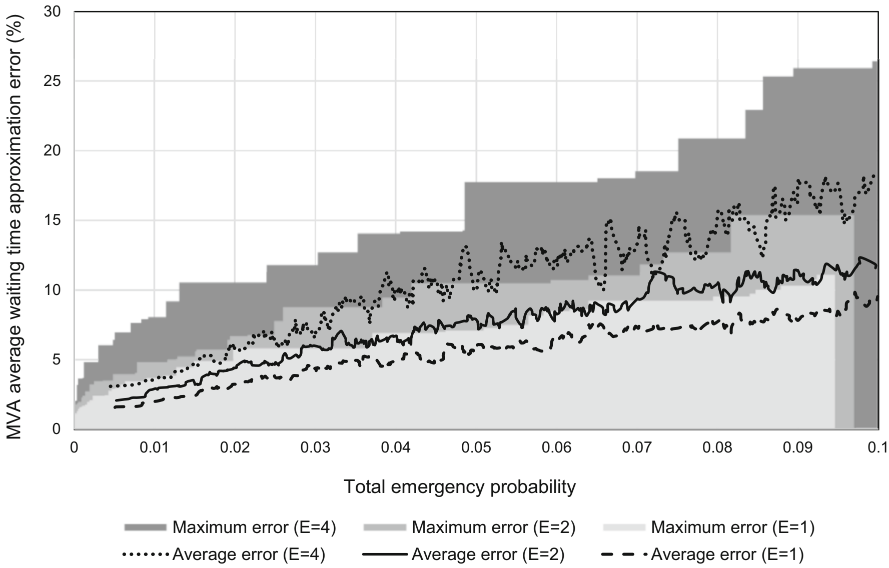

Fig. 3 Average and maximum relative error of the average waiting time approximation (MVA) in comparison with exact solutions for different number of service engineers and as a function of the total emergency probability (2250 instances, $K = 5$)

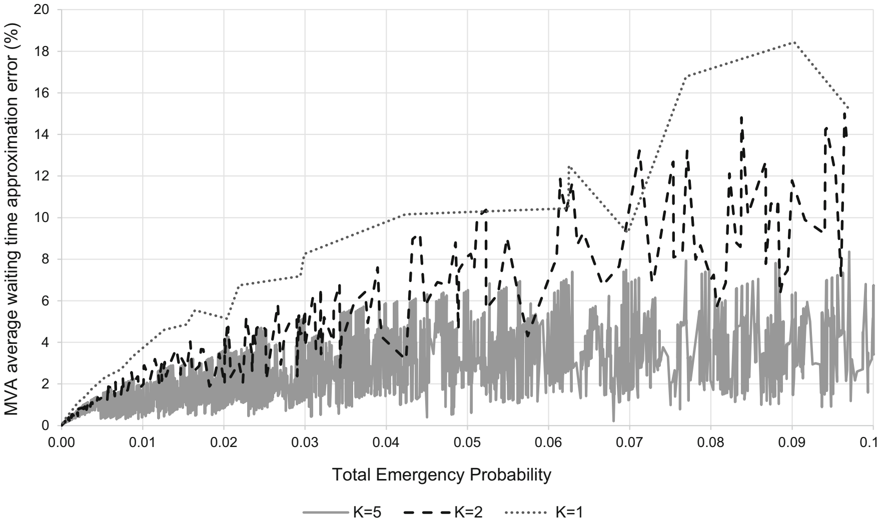

Fig. 4 Approximate average waiting times error for problems with one, two, and five type of spare parts (17300 instances, $E = 1$)

Moreover, the MVA approximation for the average waiting time works best for a single server queue ($E = 1$), and increasing the number of service engineers decreases the quality of our approximation. Here, we just examine the total emergency probability and the number of engineers as they seem to be the most influential factors on the approximation error. However, there are other parameters and factors that have an impact on the approximate average waiting time error, such as stock levels, service engineers workload, and the individual emergency probability of each spare part type. Except for the emergency probability and the number of service engineers, we can not draw a specific conclusion on the effect of other parameters on the approximation error. For example, it seems that the MVA approximation method works better when the stock level is higher, for the same values of the emergency probability and the number of service engineers. However, we have found some instances for which this does not hold.

To see how the number of spare part types affects the approximation error, we tested the model for instances with one, two, and five types of spare parts. As shown in Fig. 4, for the same value of the total emergency probability, the approximation error is lower when there are more types of spare parts. The reason behind this behavior is related to the superposition of the arrival processes. We know from the Palm–Khintchine theorem that the superposition of N independent renewal processes converges to a Poisson process as N goes to infinity (cf. Heyman and Sobel 2003, Chapter 5.8). When we have many types of spare parts, the arrival process for service engineers will be the superposition of a large number of independent processes. This leads to a process that in a sense is more similar to a Poisson process. The MVA approximation method gives the exact solution when the arrival process is Poisson. Therefore, we expect that the MVA approximation method works better when there are more spare part types in the system.

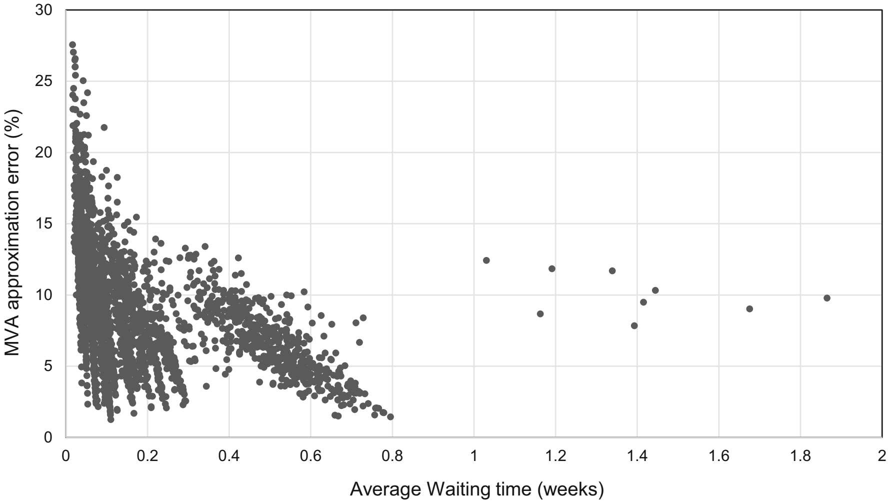

Fig. 5 MVA approximation error for different values of the average waiting time (2250 instances, $E = 1$)

The MVA approximation error versus the average waiting time value is illustrated in Fig. 5. Note that the (percentual) approximation error can be larger when the average waiting time has a low value. It generally means that we expect to have a low absolute error in all instances. In these instances, all the approximation errors that are higher than 15% are for cases where the average waiting time is less that 0.2 week. Therefore, for cases where the approximation error (in percentage) is high, the absolute error is still sufficiently low. In our instances, the highest absolute error for problems with higher than 15% error is 0.027 weeks ($\simeq$1 h).

As expected, when the emergency probability is not that low ($\geqslant$0.05) and there is a small number of spare parts, the MVA approximation method is not accurate. However, this is not a major problem as in practice the number of spare part types is high enough to have a very low approximation error. In addition, for problems with a small number of spare parts types, we can always use the exact evaluation using the matrix-geometric approach as explained in Sect. 4.1. Furthermore, we observe that the approximate average waiting time in the service engineers queue is higher than the exact ones. It means, in an approximate solution where the maximum average waiting time is satisfied, we know that the exact average waiting time is also less than the maximum waiting time.

Next, we validate the LT approximation method and compare it with the MVA approximation. We know that the MVA approximation method works well when there is a high number of spare part types and the emergency probability is low. So, we are interested to see whether the LT approximation method can be a better alternative for the cases with a medium number of spare part types ($5 < K < 50$) and relatively high emergency probability (>0.05). We compare the LT and MVA approximation methods, for instances, with two and five types of spare parts. Figure 6 shows the average approximation error for the expected waiting time as obtained by the MVA and LT evaluation methods. We use the same parameter settings as in Fig. 3. As we

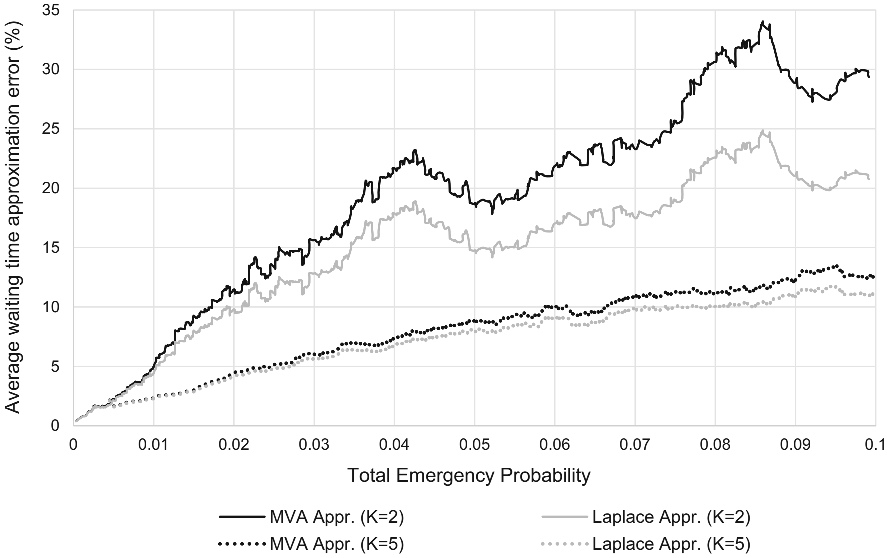

Fig. 6 Average error in expected waiting time approximation for MVA and LT evaluation methods in instances with two and five types of spare parts. (5000 instances)

expected, the LT method gives always better approximation. The average difference is larger when the total emergency probability is higher and the number of spare part types is smaller.

In all instances, we have assumed that the replenishment lead time for all spare part types is exponentially distributed. Now, we explore the sensitivity of the problem with respect to the replenishment time distribution. Therefore, we solve a problem with different replenishment time distributions (with the same rate), to see how much it affects the average waiting time value in the service engineers queue. We test a problem with two types of spare parts for six different replenishment time distributions; Erlang 2, Erlang 5, hyper-exponential, deterministic, uniform and exponential distributions (hence a wide range of values of the coefficients of variation) with the same mean value. We solve the problem for different stock levels and numbers of service engineers with all these replenishment time distributions. Although we get almost the same results (average waiting time in the service engineers queue) for Erlang and exponential distributions, using other replenishment time distributions gives different values for the average waiting time. We get up to 50% differences for the average waiting time value using hyper-exponential, deterministic, or uniform distributions in comparison with solutions where the exponential replenishment time is used. Therefore, we can say that the problem is sensitive to the replenishment time distribution with various values of the coefficients of variation.

In conclusion, we proposed one exact and two approximation methods to evaluate the model for a given policy. Note that all the instances used for our numerical validation are small enough to be evaluated using the exact evaluation method. We used these instances to show how each approximation method performs with respect to the different parameters and in comparison with other evaluation methods. In summary, the exact evaluation method is the best option for small size problem. Using this method is feasible for problems with up to 5 spare part types. Generally, in real situations, there are many different spare part types and the desired emergency probability is very small. In this case, the MVA approximation gives reliable solutions, and the approximation error for the average waiting time in the service engineers queue is sufficiently small. However, for cases where the emergency probability is not that small or there is a limited number of spare part types, the LT approximation method gives more accurate results. The LT approximation method is not as fast as the MVA approach, but the approximation error is smaller, especially for smaller problems or for problems with a relatively high emergency probability. The LT method is not computationally expensive and can be used for problems with any number of spare part types. However, for large problems ($K > 50$) where the approximation error is expected to be close to zero and the difference between LT and MVA approximation error is negligible, we recommend to use the MVA method as it is faster than the LT approximation.

For the numerical comparison, we only used instances with a small number of spare parts ($K \leqslant 5$). In our experiments, we measured computation times in milliseconds, and they were mostly zero. However, for the exact evaluation, some instances (where $K = 5$ and $S_k \geqslant 4, \forall k = 1, \ldots, 5$) required almost an hour. So, for these instances, there is a huge difference in computation time between the exact and the approximate evaluation. This difference will even be larger for instances with a larger number of spare parts, or higher values of base-stock levels, and a larger number of service engineers, see Eq. (10). When the number of spare parts becomes too large, the exact evaluation will be impossible due to the size of the state space.

5 Optimization problem

In the previous section, we described how the system under a given policy, i.e., a choice for all spare parts base-stock levels and for the number of service engineers, can be evaluated in an approximate way (for large size problems) and in an exact way (for small size problems). In this section, we find a suboptimal policy by minimizing the total average costs under total average waiting time constraints. We consider the following cost factors and parameters:

C_k^l: Cost of type-k repair call emergency shipment
H_k: Holding cost per item per unit of time for a type-k spare part (this cost applies to parts in the inventory and the pipeline)
O: Cost of hiring a service engineer per unit of time
$W^{\max}$: Maximum accepted average waiting time

For the objective function, since there is a cost for lost calls (emergency shipments), besides the number of engineers and stock levels, the emergency rate for each spare part type is needed. Note that the emergency rate of type-k spare part is a nonlinear but convex function of the stock level, see (12). For the whole system, there is a maximum average waiting time that must be satisfied. So, there exists a constraint on the total average waiting time, given in (37). The total average waiting time is a function of the number of service engineers E and all spare part stock levels S_k, $k = 1, \ldots, K$.

Let $\mathrm{S} = \{S_1, \ldots, S_K\}$ be the vector of base-stock levels. The optimization problem is then formulated as follows:

$$\min_{\mathrm{S},E} TC(\mathrm{S}, E) = O.E + \sum_{k=1}^{K} H_k S_k + \sum_{k=1}^{K} C_k^L \lambda_k^L(S_k) \tag{38}$$

$$W(\mathrm{S}, E) \leq W^{\max}, \tag{39}$$

where $W(\mathrm{S}, E)$, as given in Eq. (37), is the sum of the expected waiting times for the emergency shipments and the service engineers, and $\lambda_k^L(S_k)$ is the emergency rate given in (12).

5.1 Optimization algorithm

The optimization problem is an integer programming problem with a nonlinear objective function and constraint. We provide a greedy heuristic algorithm with local search in which we use our evaluation methods to determine a near-optimal policy to minimize the total average costs under the maximum average waiting time constraint. Similar greedy methods are used for spare parts inventory models with lateral transshipment. Wong et al. (2005) have shown that the greedy algorithm followed by local search performs very well for their multi-item multi-location spare parts systems problem. Using a greedy algorithm (without a local search) in Van Kranenburg and Houtum (2009) for a partial pooling model in spare part networks also gives reasonable results (compared to the Dantzig–Wolfe lower bound).

The emergency probability, defined in (11), is a decreasing and convex function in the number of servers S_k. So, the emergency rate for each part type is a decreasing and convex function of its stock level. Therefore, without considering the waiting time constraint, we can minimize the holding and the emergency costs for each spare part type separately with a simple greedy search. Suppose S^0 is the vector solution of this minimization. Given S^0, we find the minimum number of service engineers, E^0, such that the service engineers queue workload is less than one (stable queue). To solve the main problem, we start with (S^0, E^0) and follow the search algorithm outlined below. Note that (S^0, E^0) may be an infeasible solution (does not satisfy the average waiting time constraint). In this greedy heuristic, we first find a feasible solution. Then, in the last step, we attempt to improve the solution, using local search, by changing the solution while it remains feasible.

The average waiting time for the emergency shipments (W^S) is decreasing in stock levels while the average waiting time in the service engineers queue (W^E) is an increasing function in stock levels. So, $W(\mathrm{S}, E)$ is not generally a monotone function in S. Therefore, both decreasing and increasing spare parts stock levels may increase the total cost and decrease the total average waiting time. This makes the greedy algorithm more challenging and different from common greedy methods that are used in spare parts inventory problems. For the service engineers, increasing the number of engineers always increases the total cost and decreases the total average waiting time. With these observations, we now present the greedy search algorithm.

1. Start with solution $\mathrm{S} = \mathrm{S}^0$ (minimization of holding and emergency costs solely) and $E = E^0 = \lceil \frac{\gamma}{\eta} \rceil$.
2. Given S^0 and E^0 found in Step 1, calculate the total average waiting time. If it is less than $W^{\max}$ go to Step 5, otherwise, go to the next step.
3. Calculate Δ, Δ^E, and for each type-k spare part, Δ_k^+ and Δ_k^- using the formulas below.

$$\Delta^E = \frac{W(\mathrm{S}, E) - W(\mathrm{S}, E+1)}{TC(\mathrm{S}, E+1) - TC(\mathrm{S}, E)} = \frac{W(\mathrm{S}, E) - W(\mathrm{S}, E+1)}{O},$$
$$\Delta_k^+ = \frac{W(\mathrm{S}, E) - W(\mathrm{S} + e_k, E)}{\max\{\epsilon, TC(\mathrm{S} + e_k, E) - TC(\mathrm{S}, E)\}},$$
$$\Delta_k^- = \frac{W(\mathrm{S}, E) - W(\mathrm{S} - e_k, E)}{\max\{\epsilon, TC(\mathrm{S} - e_k, E) - TC(\mathrm{S}, E)\}},$$
$$\Delta = \max{}_k \left\{\Delta^E, \Delta_k^+, \Delta_k^-\right\},$$

where ϵ is a very small positive number, e_k is a basis vector with its kth element equals to 1 and all other elements equal to 0. Note that, if Δ equals Δ_E, we increase the number of service engineers by one. Otherwise, if Δ equals Δ_k^+, we increase the type-k spare part stock level by one, and if it equals to Δ_k^- we decrease it by one.
4. Calculate the total average waiting time with the updated solution. If it is less than $W^{\max}$, go to the next step. Otherwise, go to Step 3.
5. Perform a local search to decrease the total cost while the solution remains feasible. The last solution is the (sub)optimal solution.

Since we deal with an integer optimization problem, applying a local search in the last step of the algorithm may improve the solution considerably. We perform a local search in the following directions:

- Decrease the number of service engineers by one: Decreasing the number of service engineers decreases the total cost for sure. So, we decrease the number of service engineers by one, if by doing so the solution remains feasible.
- Decreasing or increasing a spare part's stock level by one: Find a spare part for which decreasing or increasing its stock level by one gives a feasible solution with a lower total cost.
- Decreasing or increasing a spare part's stock level by one and at the same time increasing or decreasing the number of service engineers by one: In this case, we combine the two aforementioned local search directions by changing a spare part's stock level and the number of service engineers by one. If as a result the total cost is decreased but the solution remains feasible, this update is sorted as the new (improved) solution.

We search for a local improvement in one of these three directions until no further improvement is possible. In each step, if there is more than one possible local improvement, start with the one that results in the highest decrease in the total cost.

In Step 3, we update the capacity of a resource that leads to the highest decrease in the total average waiting time per unit of cost. When updating a capacity decreases both the total average waiting time and the total cost, it becomes a super candidate for this step. So, we use ϵ to ensure that the denominator is positive even when the change in the total cost is negative for a resource and also to make the corresponding Δ large enough to be the best candidate for this step.

This algorithm always stops after a finite number of steps. First, the total average waiting time converges to zero, when the stock levels and the number of service engineers go to infinity. Hence, we always reach a feasible solution. Second, the case in which both the total cost and the total average waiting time decrease by changing a stock level cannot happen an infinite number of times. The total cost function has an increasing tail in stock levels, since the emergency cost converges to zero when the stock levels go to infinity.

Note that there is no guarantee that, by using this greedy heuristic, we obtain the optimal solution. Since we have an integer programming problem with a nonlinear constraint, we may have multiple local optimal solutions. This heuristic algorithm gives one of the local optimal solutions, which may be far from the global optimal solution. Therefore, to validate our optimization algorithm, we test the results of the algorithm numerically against other algorithms in the next section. We show that, although our integrated optimization algorithm is suboptimal, it performs well compared to other algorithms in terms of total cost and runtime, both on average and in most of the cases.

5.2 Numerical evaluation of the optimization algorithm

In this section, we show how using the approximate evaluation methods affects the optimization result. Furthermore, we compare the greedy heuristic with two other optimization algorithms. We illustrate how the integrated optimization model performs when compared with a separated optimization problem, where spare parts inventory and service engineers planning are optimized separately, and when compared with the genetic algorithm (optimization package of MATLAB R2014b). Finally, we analyze a case study using real data obtained from a company.

5.2.1 Approximate evaluation in optimization

We use instances with $K = 2$ (1000 instances) and $K = 5$ (150 instances) where K is the number of spare part types. The parameter settings are given in the first two columns of Table 1. For these settings, we solve the optimization problem using exact, MVA, and LT evaluation methods. Then, we compare the (sub)optimal total cost for solutions with exact and approximate evaluations (total cost error). The total cost function for given input values of stock levels (S) and number of service engineers (E) is exact. Note, with all these methods, there is no guarantee of optimality. Table 1 summarizes the result.

The runtime speed ratio shows how much faster on average we can solve the problem if we use MVA or LT approximation methods instead of the exact evaluation. Although in some instances the total cost error can be considerable, the average error for both

Table 1 Maximum and average total cost error of the (sub)optimal results using MVA and LT approximation methods for instances with 2 and 5 spare part types

Settings	K		MVA Appr.	LT Appr.
$W^{\max}$: 0.002 − 0.05week O : 700 − 2000€/week H_k : 500 − 3000€/week C_k^L : 0 − 20000€/week λ_k : 0.5 − 3/week ν_k : 0.4 − 3/week ν_k^{em} : 10 − 30/week μ_k : 3 − 20/week	2	Average total cost error (%)	0.956	0.745
		Maximum total cost error (%)	79.8	62.7
		Instances with positive error (%)	10.10	6.90
		Instances with negative error (%)	0.50	0.20
		Runtime speed ratio	4	3
	5	Average total cost error (%)	0.560	0.382
		Maximum total cost error (%)	19.00	19.00
		Instances with positive error (%)	22.63	17.52
		Instances with negative error (%)	0.73	0.73
		Runtime speed ratio	15000	9000

The total cost error is calculated as $100 \times \frac{\text{Total cost (Appr. evaluation)} - \text{Total cost (exact evaluation)}}{\text{Total cost (exact evaluation)}}$.
Positive (negative) error shows we end up with a worse (better) solution using approximate evaluation methods

approximation methods is very low (less than 1%). As given in Table 1, the average and maximum total cost error decrease considerably when we have more types of spare parts as is the case in real systems. In addition, runtime differences for exact and approximation methods increase exponentially with K. Even for problems with five types of spare parts (a rather small size problem), we can solve the optimization problem on average up to 15000 times faster by using approximate evaluation methods. In real cases, the number of spare parts is rather high, so the approximation error is expected to be close to zero and the runtime difference for exact and approximate evaluation methods would be huge.

Let us discuss the positive and negative error results. As we observe in all instances, both MVA and LT approximation methods overestimate the average waiting time. However, since there may be several local optimal points, solving the optimization problem with approximate evaluations may give a different local optimal point than the one we find using the exact evaluation method. Therefore, although on average the total cost of solutions where the approximate evaluation is used is higher than in case we use the exact evaluation, there are some cases where we get better solutions (lower total cost) by using MVA or LT approximation methods. We have found some cases where by using MVA or LT method we obtain solutions with the total cost up to 16% lower than solutions based upon the exact evaluation. In Table 1, the percentage of instances where we get worse solutions using approximate evaluations (positive error) and percentage of instances where we obtain better solutions (negative error) are given. Note that in the majority of cases we find the same total suboptimal cost using approximations and the exact evaluation (but in a much faster runtime). Furthermore, the optimization solutions that we obtain by using the approximate evaluation is a feasible solution for the real problem, since the approximate average waiting time is an upper bound of the exact value. In summary, we obtain a reliable feasible solution in a very efficient (fast) way by using the proposed approximation methods for problems with more than five types of spare parts.

5.2.2 *Optimization algorithms comparison*

Here, we use the same instances as in Table 1, to compare our integrated optimization algorithm with the separated optimization problem and the genetic algorithm(GA). GA is well known for solving nonlinear integer programming problem with nonlinear constraints. We use the exact waiting time for both integrated and separated optimization problems, and we test GA for both exact and MVA evaluation methods (slowest and fastest evaluation methods). Moreover, to compare the runtime of different algorithms, we also include our greedy algorithm with LT evaluation in this comparison. For the separated optimization problem, first we optimize the spare parts inventory by assuming that there is an unlimited number of service engineers. Then, we determine the smallest number of engineers such that the total average waiting time becomes less than the maximum acceptable one. Usually, the separated optimization converges in a smaller number of iterations than the integrated optimization. However, the (exact) evaluation of each iteration requires the same time as in the integrated optimization. There is no guarantee that the global optimal solution is obtained in any of these algorithms.

We expect that the optimal solutions of the integrated planning problem are never worse than solutions of the separated planning problem. However, we know that the greedy heuristic algorithm that we use does not necessarily yield the global optimal solution for the integrated optimization problem. Hence, we may get better solutions in separated optimization problem if the integrated optimization solution is far from the global optimum. Fortunately, this happens only in a very small number of cases as we will show below.

For these five optimization methods, we check the maximum and the average total (suboptimal) cost differences. In Table 2, we show for each optimization method, the total suboptimal cost (average and maximum) as compared with the best solution among other methods (error). Moreover, we show in what percentage of the instances each algorithm gives the best (or equal to the best) solution among the others. In addition, the normalized runtime ratio of each method is given in the table. This ratio shows how much time (on average) it takes to solve the problem with each algorithm as compared with the fastest one.

We cannot draw a specific conclusion under which condition each of these optimization algorithms performs best. In the instances with two types of spare parts, the greedy algorithm (with LT evaluation) and the separated optimization are the fastest algorithms among these five methods. Between these two, the greedy algorithm performs better on average. Although the greedy heuristics with exact evaluation are 3 times slower, we obtain on average more than 2% lower total costs. In addition, we reach the best solution in a larger number of instances using the greedy algorithm with exact evaluation. In five types spare parts instances, the greedy algorithm with LT evaluation is the fastest while other algorithms have much higher runtime ratio. However, its average and maximum error are not that much different than that of the greedy algorithm with exact evaluation. The performance of separated optimiza-

Table 2 Maximum and average total cost error for each optimization algorithm

K		Greedy (exact)	Greedy (LT)	Sep. Opt.	GA (exact)	GA (MVA)
2	Average error (%)	0.42	1.16	*2.89*	**0.00**	0.93
	Maximum error (%)	50.00	62.68	*96.89*	**2.21**	53.08
	Best Solution (%)	95.5	89.2	*84.1*	**99.9**	89.0
	Runtime ratio	3	**1**	**1**	*400*	25
5	Average error (%)	0.17	0.55	*2.96*	**0**	0.43
	Maximum error (%)	9.90	19.00	*38.94*	**0**	11.13
	Best Solution (%)	94.20	77.54	*68.12*	**100**	78.26
	Runtime ratio	3500	**1**	600	*1500000*	30

The percentage number of instances in which each algorithm gives the best solution is given. Runtime ratio shows the normalized computational complexity. The best solutions are in bold and the worst are italicized

tion is worse (in terms of error and runtime) compared to other ones. Therefore, it is unnecessary to test the separated optimization with approximate evaluations. The GA method with exact evaluation performs best but is computationally much more expensive.

In summary, as we discussed before, for problems with more than five types of spare parts, using the approximate evaluation is highly recommended. By increasing the size of the problem, the computation time ratio of the exact against the approximation evaluation increases exponentially, and at the same time, the approximation error decreases considerably (see the drop in maximum error in Table 2). The separated optimization algorithm is not an interesting option in any case. Both the greedy heuristic and GA (with approximate evaluations) are possible options to optimize the problem. The GA gives better solutions but needs more time to run. In the cases where the runtime is important, the greedy algorithm is a better option. Note that the GA is sensitive to the lower and upper bounds that we choose at the beginning. Without a good estimator, one may come up with very conservative lower and upper bounds. In this case, the GA becomes too slow and may yield worse solutions. Highly conservative bounds may cause the GA to end up in a solution far from the global optimal solution. To start with better bounds for the GA, we use the solutions of the greedy algorithm.

5.3 Case study

In this section, we perform a case study with data obtained from a real maintenance logistics problem of a company. In this problem, there are 93 different spare parts and a team of service engineers that are responsible for the repair. Most of the parameters are based on the data of the company. However, we have to estimate the emergency shipment costs and the service rates. We assume that these two parameters are the same for all types of spare parts. All these spare parts regard a single system (same downtime cost), and they are all shipped from the same location. Therefore, using the same value for emergency shipment costs is a logical choice. Moreover, replacement

Table 3 Total integrated optimization cost saving compared to the separated planning for different values of the maximum average waiting time

$W^{\max}$ (h)	Total cost saving (%)	Total emergency probability (%)	
		Separated solution	Integrated solution
6	27.7	12.00	9.04
4.5	23.6	9.00	8.64
3	21.5	6.00	5.59
1.8	13.6	3.60	3.15
0.9	18.8	1.80	1.31
0.3	14.4	0.60	0.54

The total emergency probability of integrated and separated solutions are given in each case

of each spare part type has similar complexity (same level in the product configuration tree), so the service rate is roughly the same for all types of repair jobs.

In this section, we compare the total cost for the integrated and the separated planning, and we investigate the use of approximate evaluations in the optimization algorithms for this (large size) problem. We solve this problem with the greedy heuristic algorithm and the separated optimization, using MVA and LT approximation and for different target service levels ($W^{\max}$). The results are summarized in Table 3. The coefficient of variation of the total arrival process to the service engineers queue in all solutions is almost 1 (>0.999). It means that this arrival process is almost a Poisson process (note that the correlation between the inter-arrivals is negligible). This causes the approximation error in the MVA and LT methods to be close to zero ($<0.1\%$). Therefore, we can use the approximate evaluation methods without any concern. Both methods lead to almost zero approximation error, so using MVA or LT approximation results in the same solutions (but MVA is faster).

As given in Table 3, the integrated optimization always gives a better solution. Integrated optimization of spare parts and service engineers results in up to 27% cost savings compared to the separated planning. The separated optimization always gives a higher total emergency probability.

Now, we are interested to see how changes in emergency shipment cost and replenishment rate affect the (sub)optimal solution. Suppose there are four possible supply options which the service provider can use for the emergency shipment; swift (most expensive), fast (expensive), normal and slow (cheap) shipment options. Table 4 shows the emergency shipment cost and rate for each of these options (in each case equal for all types of spare parts). Suppose the maximum accepted average waiting time is 0.9 h. We solve the problem by using each of these options to see which one results in the solution with the lowest total cost. The result is presented in Table 4. In all cases, we meet the maximum average waiting time constraint, but fast shipment option gives the lowest total cost, and therefore, it is the best choice to use for emergency shipment. The same analysis can be done for other parameters in the problem to get more managerial insights for real problems.

Table 4 Total cost comparison for different emergency shipment options with different emergency cost and replenishment rate

	Slow	Normal	Fast	Swift
C^L (€)	12000	20000	28000	48000
ν^{em} (per year)	30	60	70	80
Total cost (€)	4434698.92	4353605.08	**4293061.56**	4303716.12
Total emergency prob. (%)	1.09	1.31	1.54	1.76

The total cost of the best shipment option (fast shipment) is in bold

6 Conclusion

In this paper, we have introduced a new analytical model for integrated spare parts inventory management and service engineers planning. A service policy is considered in which backlogging is followed for service engineers when the spare part is available whereas the repair call is satisfied entirely via an emergency channel in case of a spare part stock-out. We have developed exact and approximate methods for performance evaluation of a given policy. When the number of spare part types is low ($K < 5$), it is computationally feasible to use the exact (matrix-geometric) evaluation method. The approximation methods yield more accurate results when there is a higher number of spare part types, and they are computationally far more efficient for large problems. Both MVA and LT approximation errors decrease considerably when the number of part types increases. In our instances, by increasing the number of spare part types from one to five, the approximation error decreases almost by half for the same value of the total emergency probability. The LT approximation method can be used in problems where neither the exact evaluation nor the MVA approximation method is sufficient. The LT is not as fast as the MVA method but still much more efficient than the exact evaluation method. For problems in which the number of spare parts types is not very large ($5 < K < 50$) or the emergency probability is not too small ($0.05 - 0.1$), the LT approximation method is more accurate and reliable than MVA for the performance evaluation. We may conclude that among these three evaluation methods there is a suitable one for each type of problem.

For the optimization problem, we use the evaluation methods in a fast greedy heuristic to determine close to optimal base-stock levels and number of service engineers. We have shown that the optimization problem can be solved in a much faster time by using approximate evaluations, while the total cost difference is negligible, specifically for larger problems ($K \geqslant 5$). In addition, we have compared the greedy algorithm optimization with separated optimization and with a genetic algorithm (GA). Although in problems with two and five types of spare parts, the GA gives better solutions than the greedy algorithm, it is more time-consuming. Moreover, the GA works better if we use the greedy algorithm solution to determine better lower and upper bounds for the GA. Finally, we used the greedy heuristics algorithm in a case study with 93 types of spare parts and compared it with separated optimization. In this problem, we showed that there can be up to 27% cost savings using the integrated planning of spare parts and service engineers as compared to the separated planning.

The presented model formulation and application can be extended in several ways. First, we assumed that in all repair calls a service engineer is needed to do the repair job. However, the model can simply be extended by assuming that not all repair calls need a service engineer. In this case, repair call arrival rates to the service engineers queue should be modified by multiplying them with a probability. Also, other arrival processes and replenishment and service time distributions should be studied to widen the area of model application. Furthermore, other service-level formulations, such as the percentile waiting time, are interesting to investigate. Approximate evaluation methods that are presented in this paper are appropriate and satisfying. However, one may think of other approximation methods, like the two-moment type of approximations (see, e.g., Tijms 2003). But for this approximation, we need the results of *GI/D/E* and *D/G/E* queues which are not known in closed form.

In this model, we consider a service policy that is applicable to many real situations. However, this is not the only justified scenario for such service logistics systems. Here, we assume that when the requested spare part is not available, the repair call will be satisfied entirely by an emergency channel. For future research, it would be of interest to consider a system in which the emergency channel just provides the spare parts and the internal service engineers must execute the emergency repair calls as well as the regular ones.

All in all, the presented approximate evaluation methods are very appropriate to use for applications in practice. First, the approximation error becomes negligible for real problems in which often the number of spare part types is rather high. Second, in the optimization of large problems, by using approximate evaluation methods we can find the solution much faster while the total cost error is almost zero. Furthermore, although the presented greedy heuristic algorithm does not give the optimal solution, it can result in a solution with much lower total cost in comparison with a separated optimization procedure.

7 Appendix 1: Inter-arrival times of type-*k* spare parts to the service engineers queue

In Sect. 4.2.3, we noted that the inter-arrival times of type-*k* repair calls in the service engineers queue has a phase-type distribution with the following Laplace–Stieltjes transform function.

$$\widetilde{X}(z) = \frac{\lambda_k\,(S_k\nu_k + (1-d_k)z)}{(\lambda_k + z)(S_k\nu_k + z)},$$

where

$$d_k = \frac{\pi_k(S_k - 1)}{1 - \pi_k(S_k)},$$

and $\pi_k(i), i = 1, \dots, S_k$ are the steady-state probabilities of the type-k spare parts in the pipeline (parts on-order). In this appendix, we prove this result. Depending on the spare part pipeline state, the arrival rate to the service engineers queue will be different. Note, we are interested in the state of spare part pipeline upon a failure and just before taking the spare part. When a type-k failure happens and there is at least one (type-k) part in stock (the pipeline is at most $S_k - 1$), a part is taken and the repair call arrives at the service engineers queue. Upon the failure, if there are less than type-k parts in the pipeline, at least one more part remains in the stock for the next type-k arriving call. Therefore, the inter-arrival time is exponentially distributed with rate λ_k. However, when the arriving call observes $S_k - 1$ parts of type k in the pipeline, it empties the spare parts type-k stock and increases the pipeline size to S_k, so a subsequent arrival to the service engineers queue will only occur after a replenishment with rate $S_k \nu_k$ and then the time until the next failure of type-k. So, in this case, the inter-arrival time will be the sum of two random variables exponentially distributed with rate $S_k \nu_k$ and λ_k. As long as the pipeline size has its maximum value, there is no arrival at the service engineers queue. We can model this inter-arrival time by means of a phase-type distribution with two transient states. Suppose state 1 is when there are less than $S_k - 1$ parts in the pipeline and state 2 is when there are $S_k - 1$ parts in the pipeline of the spare part inventory of type k. The matrix G_k gives the generator matrix of this phase-type distribution:

$$G_k = \begin{bmatrix} A_k & A_k^0 \\ 0 & 0 \end{bmatrix} = \left[\begin{array}{cc|c} -\lambda_k & 0 & \lambda_k \\ S_k \nu_k & -S_k \nu_k & 0 \\ \hline 0 & 0 & 0 \end{array}\right], \tag{40}$$

and the initial probability of the (absorbing) Markov chain is given by

$$D^k = \left[\frac{1-\pi_k(S_k-1)-\pi_k(S_k)}{1-\pi(S_k)} \;\; \frac{\pi_k(S_k-1)}{1-\pi(S_k)} \right] = \left[1 - d_k \;\; d_k \right]. \tag{41}$$

For a phase-type distribution, the Laplace–Stieltjes transform function of time until absorption (time between type-k arrivals to service engineers queue) is equal to (see, e.g., Neuts 1981)

$$\widetilde{X}(z) = D^k.(zI - A_k)^{-1}.A_k^0 \tag{42}$$

$$= \frac{\lambda_k \left(S_k \nu_k + (1 - d_k)z\right)}{(\lambda_k + z)(S_k \nu_k + z)}. \tag{43}$$

8 Appendix 2: Fitting a Coxian-2 distribution to a superposition of arrival processes

To obtain an approximate distribution, it is common to fit a phase-type distribution on the mean and the coefficient of variation of a given positive random variable. In our problem, since the analysis of the total arrival process to the service engineers queue is complex for a large number of spare part types, we propose to fit a phase-type distribution to the inter-arrival time distribution using its first two moments. The exact inter-arrival time mean value, $^1/_\gamma$, is given by (14) and (23). An approximate value of the squared coefficient of variation of the inter-arrival time is given in (25). As proposed by van Vuuren (2007), p. 20, in case $0.5 \leq c_a^2$, we can use a Coxian-2 distribution for a two-moment fit. Suppose θ_1 and θ_2 are the rates of first and second phase, respectively, of the Coxian-2 distribution, and let q denote the transition probability from phase one to two. Then, the parameters of the fitted Coxian-2 distribution are given by

$$\theta_1 = 2\gamma, \tag{44}$$

$$q = {}^{0.5}/_{c_a^2}, \tag{45}$$

$$\theta_2 = q\theta_1. \tag{46}$$

The Laplace transform function of a Coxian-2 distribution with parameters θ_1, θ_2, and q gives

$$\widetilde{X}(w) = \frac{\theta_1(\omega(1-q)+\theta_2)}{(\theta_1+\omega)(\theta_2+\omega)}. \tag{47}$$

Using (44–46), we find

$$\widetilde{X}(w) = \frac{\gamma\left(2\gamma+(2c_a^2-1)\omega\right)}{(\omega+2\gamma)(\omega c_a^2+\gamma)}. \tag{48}$$

9 Appendix 3: Approximate coefficient of variation of inter-arrival times of the superposition of arrival processes

In this section, we show how we find a simple and accurate approximation for the coefficient of variation of the superposition of two independent arrival streams to the service engineers queue. We do it in three steps.

First, note that the coefficient of variation of each single arrival stream in our model is always between 0.5 to 1. Therefore, Coxian-2 is a good candidate to fit to the arrival processes. As explained in "Appendix 2" where we fit a Coxian-2 distribution to the total arrival process, we can do the same for individual arrival streams, see (44–46).

Second, suppose we have two identical Coxian-2 arrival processes with parameters $\theta_1 = 2\gamma$, $q = {}^{0.5}/_{c^2}$, $\theta_2 = q\theta_1$. The distribution of an arbitrary inter-arrival time of the superposition of these two arrival processes can be described by a phase-type distribution with 3 phases, numbered 0, 1, 2. In phase i exactly i arrival processes are in the second phase of the inter-arrival time and $2-i$ arrival processes are in the first phase. The generator matrix Φ and the initial probability vector β of this phase-type distribution are as follows (for more details see van Vuuren 2007, p. 23):

$$\Phi = \begin{pmatrix} -4\gamma & {}^{2\gamma}/_{c^2} & 0 \\ 0 & -2\gamma - {}^{\gamma}/_{c^2} & {}^{\gamma}/_{c^2} \\ 0 & 0 & -{}^{2\gamma}/_{c^2} \end{pmatrix}, \tag{49}$$

$$\beta = \left({}^1/_2 \; {}^1/_2 \; 0\right). \tag{50}$$

The squared coefficient of variation of this phase-type process equals

$$c_a^2 = \frac{2\beta\Phi^{-2}e}{\left(\beta\Phi^{-1}e\right)^2} - 1 = \frac{c^2(2+c^2)}{1+2c^2}, \tag{51}$$

where e is the column vector with all elements equal to one.

Third, as an approximation, we can replace two independent arrival streams with arrival rates γ_1 and γ_2 and coefficient of variations c_1 and c_2 with two identical Coxian-2 arrival streams with an arrival rate $(\gamma_1+\gamma_2)/2$ and squared coefficient of variations L_2 which is given by

$$L_2 = \frac{\gamma_1}{\gamma_1+\gamma_2}c_1^2 + \frac{\gamma_2}{\gamma_1+\gamma_2}c_2^2. \tag{52}$$

Note that, $\gamma_1/(\gamma_1+\gamma_2)$ and $\gamma_2/(\gamma_1+\gamma_2)$ are the fraction of arrivals that are of types 1 and 2, respectively (see 16). Then, we can use Eq. (51) as an approximation for the squared coefficient of variation of the superposition process of the two arrival streams.

$$c_a^2 = \frac{L_2(2+L_2)}{1+2L_2}$$

The same method applies when we have three or more arrival streams (see van Vuuren 2007, p. 23). For three arrival steams, the squared coefficient of variation of the superposition process equals

$$c_a^2 = \frac{L_3(3+6L_3+L_3^2)}{1+5L_3+4L_3^2}.$$

where L_3 is given by Eq. (26). However, as a computationally more efficient procedure, we can use Eqs. (27) and (28) iteratively to find the coefficient of variation of the superposition process when there are more arrival streams.

References

Agnihothri SR, Karmarkar US (1992) Performance evaluation of service territories. Oper Res 40(2):355–366

Agnihothri SR, Mishra AK (2004) Cross-training decisions in field services with three job types and server-job mismatch. Decis Sci 35(2):239–257

Agnihothri SR, Mishra AK, Simmons DE (2003) Workforce cross-training decisions in field service systems with two job types. J Oper Res Soc 54(4):410–418

Akşin OZ, Harker PT (2003) Capacity sizing in the presence of a common shared resource: dimensioning an inbound call center. Eur J Oper Res 147(3):464–483

Albin SL (1984) Approximating a point process by a renewal process, ii: superposition arrival processes to queues. Oper Res 32(5):1133–1162

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

Avsar ZM, Zijm WH, Rodoplu U (2009) An approximate model for base-stock-controlled assembly systems. IIE Trans 41(3):260–274

Axsäter S (1990) Modelling emergency lateral transshipments in inventory systems. Manag Sci 36(11):1329–1338

Basten RJI, van Houtum GJ (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

Brickner C, Indrawan D, Williams D, Chakravarthy SR (2010) Simulation of a stochastic model a service system. In: Proceedings of the 2010 winter simulation conference, pp 1636–1647

Chopra S, Meindl P (2013) Supply chain management: strategy, planning, and operation. Pearson International Edition, Pearson

Dayanik S, Song JS, Xu SH (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

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

Hertz P, Cavalieri S, Finke GR, Duchi A, Schonsleben P (2014) A simulation-based decision support system for industrial field service network planning. Simul-Trans Soc Model Simul Int 90(1):69–84

Heyman DP, Sobel MJ (2003) Stochastic models in operations research: stochastic processes and operating characteristics, vol 1. Courier Corporation, North Chelmsford

Hoen KM, Güllü R, Van Houtum G, Vliegen IM (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

Koole G, Pot A (2006) An overview of routing and staffing algorithms in multi-skill contact centers. Dept of Mathematics, Vrije Universiteit, Amsterdam, pp 1–42

Kranenburg A, Van Houtum G (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

Kukreja A, Schmidt CP, Miller DM (2001) Stocking decisions for low-usage items in a multilocation inventory system. Manag Sci 47(10):1371–1383

Kutanoglu E (2008) Insights into inventory sharing in service parts logistics systems with time-based service levels. Comput Ind Eng 54(3):341–358

Latouche G, Ramaswami V (1999) Introduction to matrix analytic methods in stochastic modeling, vol 5. Society for Industrial and Applied Mathematics, Philadelphia

Lu Y, Song JS, Yao DD (2003) Order fill rate, leadtime variability, and advance demand information in an assemble-to-order system. Oper Res 51(2):292–308

Muckstadt JA (2005) Analysis and algorithms for service parts supply chains. Springer, Berlin

Neuts MF (1981) Matrix-geometric solutions in stochastic models: an algorithmic approach. Courier Corporation, North Chelmsford

Ormeci EL (2004) Dynamic admission control in a call center with one shared and two dedicated service facilities. IEEE Trans Autom Control 49(7):1157–1161

Papadopoulos HT (1996) A field service support system using a queueing network model and the priority mva algorithm. Omega-Int J Manag Sci 24(2):195–203

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

Sherbrooke CC (2004) Optimal inventory modeling of systems: multi-echelon techniques, vol 72. Springer, Berlin

Shumsky RA (2004) Approximation and analysis of a call center with flexible and specialized servers. OR Spectr 26(3):307–330

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

Song JS, Xu SH, Liu B (1999) Order-fulfillment performance measures in an assemble-to-order system with stochastic leadtimes. Oper Res 47(1):131–149

Takács L (1962) An introduction to queueing theory. Oxford University Press, New York

Tijms HC (2003) A first course in stochastic models. Wiley, Hoboken

van Houtum GJ, Kranenburg B (2015) Spare parts inventory control under system availability constraints, vol 227. Springer, Berlin

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

Visser J, Howes G (2007) A simulation technique for optimising maintenance teams for a service company. S Afr J Ind Eng 18(2):169–185

Vliegen I (2009) Integrated planning for service tools and spare parts for capital goods. PhD thesis, Technische Universiteit Eindhoven

van Vuuren M (2007) Performance analysis of manufacturing systems: queueing approximations and algorithms. PhD thesis, Technische Universiteit Eindhoven

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

Whitt W (1983) The queueing network analyzer. Bell Syst Tech J 62(9):2779–2815

Wong H, van Houtum GJ, 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 56(12):1419–1430