TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

A new approximate evaluation method for two-echelon inventory systems with emergency shipments

Citation for published version (APA):
Ozkan, E., Houtum, van, G. J. J. A. N., & Serin, Y. (2015). A new approximate evaluation method for two-echelon inventory systems with emergency shipments. *Annals of Operations Research*, *224*(1), 147-169. https://doi.org/10.1007/s10479-013-1401-9

DOI:
10.1007/s10479-013-1401-9

Document status and date:
Published: 01/01/2015

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

Please check the document version of this publication:

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

Link to publication

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

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

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 22. May. 2025

Ann Oper Res (2015) 224:147–169
DOI 10.1007/s10479-013-1401-9

A new approximate evaluation method for two-echelon inventory systems with emergency shipments

Erhun Özkan · Geert-Jan van Houtum · Yasemin Serin

Published online: 12 June 2013
© Springer Science+Business Media New York 2013

Abstract We consider the inventory control of repairable spare parts in a network consisting of a central warehouse, a central repair facility, and multiple local warehouses. Demands for spare parts occur at the local warehouses. If a local warehouse is out of stock, then an arriving demand is satisfied by an emergency shipment from the central warehouse or the central repair facility. Such emergency shipments are common practice for networks that support technical systems with high downtime costs. We develop a new evaluation method that provides accurate approximations for the key performance measures like fractions of demands supplied by the local warehouses or emergency shipments. The method can be easily incorporated in existing (greedy) heuristic optimization methods. Our method outperforms the approximate evaluation method of Muckstadt and Thomas (Manag. Sci. 26:483–494, 1980), as we show via a numerical analysis. Finally, we show that the performance of the system is virtually insensitive to the leadtime distribution of repairs at the central repair facility.

Keywords Spare parts · Two-echelon system · Emergency shipments · Approximate evaluation

1 Introduction

The management of spare parts has become an important issue in the capital goods industry. For many technical systems, downtime costs are high and, thus, failed parts need to

E. Özkan (✉)
Department of Industrial Engineering, University of Pittsburgh, Pittsburgh, PA 15261, USA
e-mail: erhunozkan@gmail.com

G.-J. van Houtum
Department of Industrial Engineering & Innovation Sciences, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
e-mail: g.j.v.houtum@tue.nl

Y. Serin
Industrial Engineering Department, Middle East Technical University, 06531 Ankara, Turkey
e-mail: serin@ie.metu.edu.tr

be replaced by spare parts as quickly as possible. Spare parts may be kept in stock in networks by the user itself, or by original equipment manufacturers (OEM) or third parties. Networks of spare parts typically consist of local warehouses within close proximity of installed systems and one or more layers of central and regional warehouses. In such networks, different types of flexibilities have been employed to react as quickly as possible to failures of technical systems. If a local warehouse is out of stock at the moment of a demand arrival, then it is possible to send a part from a neighboring local warehouse and/or directly from a higher-level warehouse. These options are denoted as lateral and emergency shipments, respectively. The options that are used or available depend on geographical factors and on arrangements that have been made with, for example, logistics service providers and external repair centers.

We consider repairable spare parts in a two-echelon system consisting of a central warehouse, a central repair facility, and multiple local warehouses. The repair facility, which is assumed to have an infinite repair capacity, supplies the central warehouse, and the central warehouse supplies the local warehouses. We assume a continuous review, one-for-one replenishment policy within the network (i.e., base stock control), a common policy in the literature of spare parts. As an illustration, we describe the supply chain of spare parts at Nedtrain, a train maintenance company in the Netherlands. Nedtrain has thousands of different repairables in its supply chain. The repairables have a wide price range; their price can reach up to tens of thousands of euros. A failure of a critical repairable causes downtime of the train until the failed part is replaced by a ready-for-use part, and downtime costs per hour are very high, which makes the availability of a critical repairable very important. When a demand arrives at a local warehouse, it is supplied by the local warehouse if there is on-hand stock available. If the demanded repairable is not available at the local warehouse, then an emergency shipment is made from the central warehouse if it has on-hand stock. If the central warehouse does not have on-hand stock, an emergency shipment must be requested from the repair facility to the local warehouse. The leadtime for an emergency shipment is shorter than a normal replenishment leadtime; therefore this supply option is costly. Managing this kind of inventory system needs quantitative models which take emergency shipments into account.

There are many studies on inventory control of spare parts. Sherbrooke (1968) introduced the METRIC (Multi-Echelon Technique for Repairable Item Control) model for two-echelon systems, without lateral and emergency shipments. Via the METRIC approach, expected backorder levels at all local warehouses can be computed under base stock control and assigned base stock levels. Sherbrooke (1968) approximates the realized replenishment leadtimes for the local warehouses by independent and deterministic leadtimes. Graves (1985) develops exact and approximate evaluation procedures for multi-echelon systems. In the approximation method, Graves (1985) fits a negative binomial distribution to the first two moments of pipeline stocks, and this approximation is shown to yield more accurate results than the METRIC approximation with respect to the expected backorder levels at the local warehouses. Rustenburg et al. (2003) generalize Graves' exact and approximate evaluation methods to multi-echelon, multi-indenture systems. Sherbrooke (1968) also develops a heuristic optimization method for the minimization of the total stock of multiple items under a constraint for the total number of backorders in the whole system. Wong et al. (2007) develop multiple heuristics for the same optimization problem but with a constraint per local warehouse. Saranga and Kumar (2006) look at the integrated optimization of spare parts stocks and the places where parts are repaired ('Level Of Repair Analysis'). Basten (2010) and Basten et al. (2009) first develop a computationally faster algorithm for the problem of Saranga and Kumar (2006), and next extend the integrated model to a more general structure for the fixed costs of required maintenance resources.

Andersson and Melchiors (2001) consider a two-echelon system, but, different from Sherbrooke (1968), they assume that backorders are not allowed at local warehouses, and demand which cannot be supplied by the local warehouse is lost. Based on the METRIC approximation of Sherbrooke (1968), they develop an accurate heuristic to determine a cost-effective base stock policy.

Muckstadt and Thomas (1980) extend the work of Sherbrooke (1968) to systems with emergency shipments from the central warehouse and central repair facility. Their focus is on the (heuristic) optimization of the base stock levels, which builds on the approximate evaluation method introduced in their paper. They also compare centralized and decentralized decision making. Hausman and Erkip (1994) improve the decentralized case of Muckstadt and Thomas (1980) and show that the performance of the improved single-echelon model differs from the multi-echelon model of Muckstadt and Thomas (1980) by 3–5 %.

Axsäter et al. (2004) consider a two-echelon inventory system in which emergency shipments are sent only from the central repair facility to the local warehouses. They assume that the emergency shipment time exceeds regular replenishment leadtimes from the central warehouse to the local warehouses. In contrast to other studies, Axsäter et al. (2004) assume that the central warehouse also receives direct customer demands, and this stream of demands has priority over the replenishment orders of the local warehouses. They use critical inventory levels at the central warehouse to differentiate between the demand streams. Axsäter et al. (2004) also develops a heuristic optimization method.

Alvarez and van der Heijden (2011) also consider a two-echelon inventory system but in their model only the external supplier (which is equivalent to the central repair facility in our model) is allowed to make emergency shipments. They assume that the emergency shipment from the external supplier takes more time than a shipment from the central warehouse. They derive an accurate approximate evaluation procedure for the following two performance measures: (i) the fraction of demands supplied by the local warehouses, central warehouse, or spare parts which are in the transit pipeline between the central warehouse and the local warehouse; (ii) the fraction of demands supplied by the emergency shipments from the external supplier.

Axsäter (1990) develops an approximate evaluation method for two-echelon systems with lateral shipments. Alfredsson and Verrijdt (1999) consider a two-echelon system with both lateral and emergency shipments. In case of a demand arrival, if the local warehouse is out of stock, they first check other local warehouses for a lateral shipment, then they check the central warehouse for an emergency shipment, and lastly they make an emergency shipment from the repair facility, if needed. Because of the lateral shipments, which are possible between all pairs of local warehouses (full pooling), they can aggregate all stocks in the local warehouses to calculate the fractions of demands satisfied by emergency shipments from the central warehouse and repair facility. For the latter step, they make use of a two-dimensional Markov process with respect to the central and local stock, and numerically compute the limiting distribution of this Markov process. Consequently, their approximation method is very time-consuming even for medium high base stock levels. Alfredsson and Verrijdt (1999) also execute a sensitivity analysis with respect to the distribution of the leadtime of the repair facility and the distribution of the transportation times between the central warehouse and the local warehouses. They find that the performance parameters are virtually insensitive to these distribution types.

Grahovac and Chakravarty (2001) consider the same system as Alfredsson and Verrijdt (1999), but without the possibility of emergency shipments from the repair facility (and thus with the possibility of backordering at the local warehouses). A second difference is that in case of a demand arrival, if the local warehouse is out of stock, they first check the central

warehouse for an emergency shipment, and then they check other local warehouses for a lateral shipment. Lastly, they consider emergency trigger inventory levels at the local warehouses, i.e, they allow lateral shipments not only when there is a stock-out situation, but at arbitrarily chosen levels of on-hand stock. They use a similar iterative solution methodology as Axsäter (1990). They also show that sharing of stock (via the emergency and lateral shipments) often, but not always, reduces overall system costs. Moreover, the optimal emergency trigger inventory levels are found to be -1 in most of the cases, implying that anticipation of future demand is often not beneficial.

Wong et al. (2005) develop a heuristic optimization method for a single-echelon, multi-location, multi-item system with lateral and emergency shipments. The emergency shipments can be done from a central warehouse which is assumed to have unlimited stock. The heuristic optimization is built on exact evaluations via Markov processes. Kranenburg and van Houtum (2009) consider the same system as Wong et al. (2005) but with a form of partial pooling instead of full pooling. In their system, only a limited number of main local warehouses are allowed to provide lateral shipments. They develop an approximate evaluation method in which demands for lateral shipments are modeled as Poisson overflow processes (in the spirit of the model by Axsäter 1990). In addition, they develop an efficient greedy heuristic for the minimization of total inventory, and lateral and emergency shipment costs, subject to mean waiting time constraints at the local warehouses. They show that using only some of the local warehouses as lateral shipment sources is sufficient to obtain most of the benefits of full pooling. Their work has been implemented at ASML, a manufacturer of lithography machines for the production of semiconductors. There are many more studies related to lateral shipments. For an overview, see Paterson et al. (2011).

In some networks, the use of emergency shipments is strongly preferred over lateral shipments, because lateral shipment may be more expensive, e.g. the local warehouses can be geographically dispersed and/or procedures for lateral shipments may not be well organized. Or, lateral shipments are even excluded by ensuring that the repair facility can always provide an emergency shipment. In the situation of Nedtrain, lateral shipments are not completely excluded, but they are seen as undesirable exceptions and, thus, they are excluded for inventory planning at the tactical level. Surprisingly, such networks with emergency shipments, but without lateral shipments, have scarcely been studied in the literature. To our knowledge, only the work of Muckstadt and Thomas (1980) considers supply networks of this kind.

In this paper, we introduce a new approximate evaluation method for two-echelon systems with emergency shipments but without lateral shipments. It will be shown that our method performs significantly better than the approximate evaluation method of Muckstadt and Thomas (1980). Our method is accurate and fast, and thus can well be used in greedy heuristic optimization methods for the multi-item version of our model. Such greedy optimization methods can be used for the minimization of inventory holding, and emergency and lateral shipment costs, subject to aggregate waiting time constraints per local warehouse. They have been shown to work very well for closely related systems; see Wong et al. (2005), Wong et al. (2007), and Kranenburg and van Houtum (2009). Our numerical results indicate that the performance of our system is virtually insensitive to the distribution of repair leadtimes at the central repair facility, which implies that our method works well for generally distributed repair leadtimes.

Exact evaluation for the system analyzed in this paper is possible via Markov methods if repair leadtimes and transportation times from the central warehouse to the local warehouses are exponentially distributed. But that would require a numerical solution of multi-dimensional Markov processes and then we would obtain long computation times for already medium high base stock levels.

Although we use the terminology of repairable spare parts in this paper, our model applies more generally. Consumable spare parts fit equally well into the same framework, as the central repair facility can be viewed as an external supplier.

The remainder of the paper is organized as follows. In Sect. 2, we describe our model. In Sect. 3, we describe our approximate evaluation method and we summarize the method of Muckstadt and Thomas (1980). Next, in Sect. 4, we test our method via a numerical analysis and perform a sensitivity analysis with respect to the repair lead time distribution and a cost optimization experiment. Some concluding remarks are provided in Sect. 5.

2 Model description

Consider a single-item, two-echelon inventory model with one central warehouse (CW), denoted by index 0, and N ($N \geq 1$) local warehouses (LW). Let $\mathcal{N} = \{1, 2, \ldots, N\}$ be the set of local warehouses. In addition, there is a central repair facility to which all failed parts are returned and repaired.

Demands for spare parts occur at the local warehouses. We assume that demands at local warehouse n arrive according to a Poisson process with a constant rate m_n ($m_n > 0$). Each demand at a local warehouse n stems from a failure of a part in a technical system. For each demand, one of the following procedures is applied (see also Fig. 1):

1. If local warehouse n has a part in stock, then it satisfies the demand itself. In this case, there is no delay in satisfying the demand. The failed part is sent to the repair facility. Further, the local warehouse places a replenishment order for one ready-for-use part at the central warehouse, and the central warehouse places an order for one unit at the repair facility.
2. If local warehouse n is out of stock, and there is at least one part in stock at the central warehouse, then the demand is satisfied from the central warehouse. In this case, the part is delivered via a fast emergency shipment, which leads to a delay in satisfying the demand of on average t_n^{CW} time units. The failed part is sent to the repair facility, and at the same time the central warehouse places an order for one ready-for-use unit at the repair facility.
3. If both local warehouse n and the central warehouse is out of stock, then a part is delivered from the central repair facility. We assume that the repair facility can always provide a spare part, e.g., it may finish the repair of one of the parts in the repair shop via an emergency procedure. This leads to an average delay of t_n^{RF} time units. The failed part is sent to the repair facility.

Under these procedures, the inventory position remains at a constant level at each of the warehouses. Let S_n be the constant level for warehouse n, $n \in \mathcal{N} \cup \{0\}$. Equivalently, we may say that the inventory is controlled by a base stock policy, and S_n is the base stock level at warehouse n.

The replenishment leadtime for local warehouse n is assumed to be deterministic and denoted by t_n. Obviously, replenishments are delayed when the central warehouse is out of stock. The central repair facility is assumed to follow a given planned leadtime, denoted by t_0. This implies that every order for a ready-for-use part placed by the central warehouse will be delivered after exactly t_0 time units. This is equivalent to modeling the repair facility as an ample server with deterministic service times t_0.

The main performance measures that need to be determined are directly related to the demand streams at the local warehouses. For the demand stream at local warehouse $n \in \mathcal{N}$, our aim is to approximate:

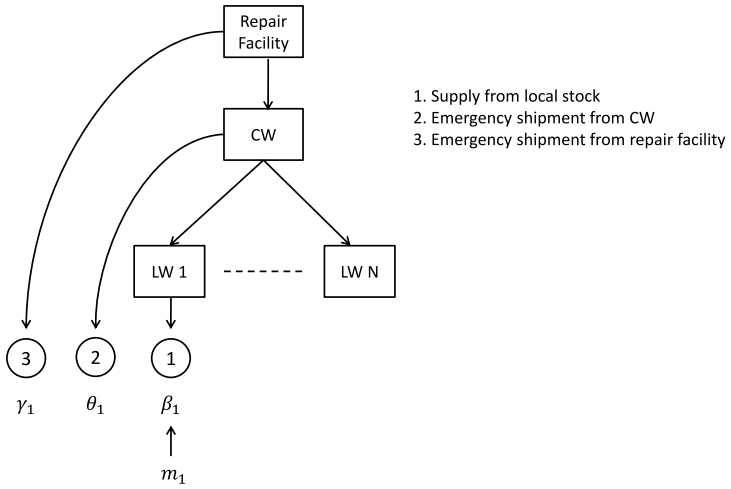


Fig. 1 Graphical depiction of the demand fulfillment process

- β_n: the steady-state fraction of demands occurring at local warehouse n that are satisfied by local warehouse n itself. This measure is also denoted as the *fill rate* of local warehouse n;
- θ_n: the steady-state fraction of demands occurring at local warehouse n that are directly satisfied by the central warehouse;
- γ_n: the steady-state fraction of demands occurring at local warehouse n that are directly satisfied by the central repair facility.

Notice that,

$$\beta_n + \theta_n + \gamma_n = 1, \quad \forall n \in \mathcal{N}. \tag{1}$$

The fractions β_n, θ_n, and γ_n are visualized in Fig. 1.

The remainder of the paper focuses on the (approximate) evaluation of the fractions β_n, θ_n, and γ_n. In optimization problems, one often minimizes a cost function subject to constraints related to downtime or availability of the supported technical systems. For example, a constraint on the mean waiting/delay time (W_n) until demands at local warehouse n are fulfilled, is common; here,

$$W_n = \theta_n t_n^{CW} + \gamma_n t_n^{RF}. \tag{2}$$

A typical total cost function would consists of inventory holding costs (for all parts in stock and in repair or in transport from the central warehouse to a local warehouse) and extra costs for demands fulfilled from the central warehouse and the repair facility:

$$C = h \sum_{n=0}^{N} S_n + \sum_{n=1}^{N} m_n \left(\theta_n C_n^{CW} + \gamma_n C_n^{RF}\right), \tag{3}$$

where h represents the inventory holding cost per stock keeping unit per time unit, C_n^{CW} represents the cost for an emergency shipment from the central warehouse to local warehouse n, and C_n^{RF} represents the cost for an emergency shipment from the repair facility to local warehouse n. For both the W_n and C, extended expressions are obtained when one wants to

optimize over multiple items. As we see, quantities such as the mean waiting times W_n and total costs C are easily obtained from the β_n, θ_n, and γ_n.

3 Solution procedures

In this section, we describe a new approximate evaluation method and summarize the approximation scheme of Muckstadt and Thomas (1980).

3.1 Approximate evaluation method

Our approximate evaluation procedure starts with a solution procedure that iteratively calculates the fill rates β_n at the local warehouses and the expected delay at the central warehouse. In each iteration, first the fill rates β_n are calculated under a given delay at the central warehouse, and next the expected delay at the central warehouse is calculated using the fill rates β_n. Below, we first describe these two steps in detail, and then we summarize the iterative procedure. Finally, we give the approximations for the fractions θ_n and γ_n.

3.1.1 Calculating the fill rates

The replenishment leadtime of local warehouse n is given by a deterministic value t_n. This time may be seen as the planned leadtime. When a replenishment order is placed at the central warehouse, its fulfillment will be delayed, and thus the realized leadtime is longer. Let W_0 be the mean delay for an arbitrary replenishment order at the central warehouse. Notice that the replenishment orders from different local warehouses experience statistically the same delays. Let LT_n denote the mean realized replenishment leadtime for local warehouse n. Then,

$$LT_n = t_n + W_0. \tag{4}$$

These realized leadtimes depend on the on-hand stock distribution at the central warehouse. The higher the basestock level at the central warehouse, the shorter the mean delay W_0. And, higher basestock levels at the local warehouses have a decreasing effect on the stream of requests for emergency shipments at the central warehouse and, thus, also a decreasing effect on the stream of emergency shipments from the repair facility (i.e., more demand has to be satisfied by the central warehouse), which then may lead to a slightly longer mean delay. In our analysis, all basestock levels are given, but the basestock levels at the local warehouses are correlated with the fill rates β_n and, thus, W_0 and β_n are dependent. For the initial computation of the β_n, we assume a zero delay, i.e., $W_0 = 0$.

The fill rates β_n are computed per local warehouse $n \in \mathcal{N}$. We assume that the realized leadtimes for replenishment orders at local warehouse n are independent and identically distributed (so this is an approximate step). Demands arrive according to a Poisson process with rate m_n. Because of the emergency shipments from the central warehouse and the repair facility, there is no backordering of demand. From the perspective of the local warehouse n, demand that is not satisfied from stock can be seen as lost demand. This implies that the local warehouse n behaves the same as an Erlang loss system (i.e., an $M/G/c/c$ queue; see e.g. Tijms 2003). Each unit of stock may be seen as a server that is occupied for on average LT_n time units when it serves a demand. In fact, the steady-state behavior of the number of outstanding replenishment orders (= S_n minus the on-hand stock) is identical to the steady-state behavior of the number of occupied servers in an Erlang loss system with S_n servers,

arrival rate m_n, and mean service time LT_n. As a result, the fill rate β_n may be obtained as the proportion of arriving customers who find an available server in the Erlang loss system.

For a general Erlang loss system with c servers and offered load ρ (the product of the arrival rate and the mean service time of a single server), let $L(c,\rho)$ denote the Erlang loss probability (i.e., the proportion of customers that are not served). It is well known that (cf. Tijms 2003)

$$L(c,\rho) = \frac{\frac{\rho^c}{c!}}{\sum_{x=0}^{c} \frac{\rho^x}{x!}}, \quad \rho > 0. \tag{5}$$

The fill rate at local warehouse n is then obtained by

$$\beta_n = 1 - L(S_n, m_n \cdot LT_n). \tag{6}$$

3.1.2 Calculating the expected delay in the central warehouse

Suppose now that the fill rates β_n are known. We want to estimate the mean delay W_0 at the central warehouse.

We model the process for the inventory level at the central warehouse as a continuous-time birth-death process (cf. Tijms 2003). Notice that the inventory level is equal to the on-hand stock minus the backordered replenishment orders from the local warehouses. Per local warehouse n, there is a demand stream of replenishment orders and a demand stream for emergency shipments. The first demand stream has rate $m_n\beta_n$ and is assumed to be a Poisson process (that this process is a Poisson process is an approximation). Demands from this stream are immediately satisfied if the central warehouse has at least one part on stock (i.e., a strictly positive inventory level) and otherwise they are backordered. The second stream has rate $m_n(1-\beta_n)$ and is also assumed to be a Poisson process. Demands from this stream are immediately satisfied if the central warehouse has at least one part in stock and otherwise they are lost (i.e., they will be satisfied by the repair facility via an emergency shipment). All demand streams are assumed to be mutually independent and independent of the actual inventory level at the central warehouse. As a result of these assumptions, the total demand stream at the central warehouse is a Poisson process with rate

$$\sum_{n\in\mathcal{N}} \big(m_n\beta_n + m_n(1-\beta_n)\big) = \sum_{n\in\mathcal{N}} m_n = m_0$$

when the inventory level at the central warehouse is strictly positive, and it is a Poisson process with rate

$$m_0' = \sum_{n\in\mathcal{N}} m_n\beta_n \tag{7}$$

when the inventory level is zero or strictly negative.

The second approximation that we make is that the deterministic leadtime t_0 at the central warehouse is replaced by an exponential leadtime with the same mean, i.e., by exponential times with rate $\mu_0 = 1/t_0$. It will be shown that the steady-state behavior of the whole system is virtually insensitive to the probability distribution of repair leadtimes; see Sect. 4.2. As noted in Sect. 1, a similar insensitivity property was also observed by Alfredsson and Verrijdt (1999) for their two-echelon system with lateral and emergency shipments. Hence, this approximation should not yield excessive errors in the estimation of the performance measures, and it facilitates that the inventory level process can be modeled as a birth-death process.

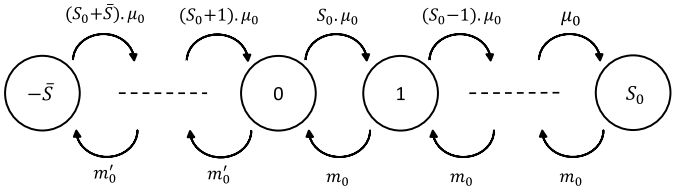


Fig. 2 Transition diagram for the inventory level at the central warehouse

Finally, we truncate the state space of the birth-death process. Because backorders can only occur when replenishment orders are placed, and the number of outstanding replenishment orders at local warehouse n can never be more than S_n, the number of backorders at the central warehouse cannot exceed $\bar{S} = \sum_{n\in\mathcal{N}} S_n$. Hence, we truncate the states x with $x < -\bar{S}$. This completes the construction of the birth-death process of the inventory level at the central warehouse; the resulting process is depicted in Fig. 2.

The mean delay W_0 is obtained from the steady-state distribution of the birth-death process. Let the steady-state distribution be denoted by $\{\pi_x\}$. The steady-state probabilities satisfy the balance equations

$$\pi_x = \begin{cases} \frac{m_0'}{(S_0-x)\mu_0} \cdot \pi_{x+1}, & -\bar{S} \le x < 0, \\ \frac{m_0}{(S_0-x)\mu_0} \cdot \pi_{x+1}, & 0 \le x < S_0. \end{cases} \tag{8}$$

By these equations, they can all be expressed as a function of π_{S_0}, and π_{S_0} itself follows from normalization. Next, the mean number of backordered demands, B_0, follows from

$$B_0 = \sum_{x=-\bar{S}}^{-1} (-x)\pi_x, \tag{9}$$

and, by Little's Law (cf. Tijms 2003), we find (notice that the rate for the total stream of replenishment orders is m_0')

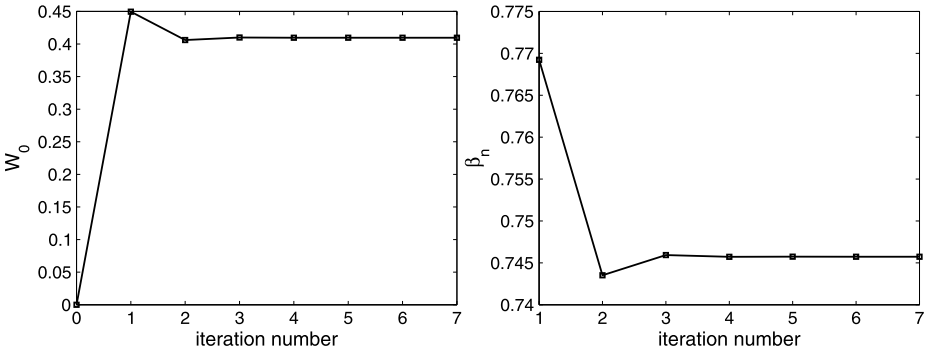
$$W_0 = \frac{B_0}{m_0'}. \tag{10}$$

3.1.3 Iterative algorithm for the approximation method

We obtain the following iterative algorithm for the computation of the fill rates β_n, $n \in \mathcal{N}$, and the mean delay W_0:

Step 0 $W_0 := 0$.
Step 1 Compute β_n via (4) and (6), $\forall n \in N$.
Step 2 Compute W_0 via (7), (8), (9), and (10).
Step 3 Repeat *Step 1* and *Step 2* until W_0 does not change more than ϵ.

With respect to the convergence of this algorithm, we have no theoretical results, but we obtained convergence for all instances used in our numerical study. The setup of the numerical study and the outcomes are reported in Sect. 4. Figures 3a and 3b show convergence of W_0 and β_n for instance 62 of the symmetric instances, where $N = 20$, $m_n = 0.1$ demands per day for all $n \in \mathcal{N}$, $t_0 = 20$ days, $t_n = 3$ days for all $n \in \mathcal{N}$, $S_0 = 40$, $S_n = 1$ for all $n \in \mathcal{N}$.



(a) W_0 at each iteration in instance 62 (b) β_n at each iteration in instance 62

Fig. 3 Convergence of W_0 and β_n in the instance 62 of symmetric instances

As we see in Fig. 3a, the initial value of W_0 is 0. In this case the corresponding β_n becomes the largest because the lower the expected delay in the central warehouse, the lower the lead times and the higher the fill rate at each local warehouse. In the following iteration, W_0 becomes the largest, because the fill rates β_n are largest and the higher the fill rates, the higher the number of replenishment orders from the central warehouse and the higher the delay for these orders. Afterwards, β_n becomes the lowest as seen in iteration number 2 in Fig. 3b. Then, W_0 becomes the second lowest, and so on. At each even numbered iteration, the β_n and W_0 decrease, and at each odd numbered iteration, the β_n and W_0 increase. At each iteration, the differences of the values for the β_n and W_0 with the values of the previous iteration decrease, and we obtain convergence.

The algorithm is robust with respect to the initial value of W_0. We experimented with different starting values, and for all possible initial values of W_0, we obtained convergence.

3.1.4 Calculation of the θ_n and γ_n

We finally approximate the fractions of demands satisfied by an emergency shipment from the central warehouse and the repair facility, respectively. Let IL_n and IL_0 be random variables which denote the inventory level in local warehouse n and the central warehouse, respectively. Then, it holds that

$$\theta_n = \mathbf{P}(IL_0 > 0, IL_n = 0).$$

By conditioning to "$IL_0 > 0$", we obtain

$$\theta_n = \mathbf{P}(IL_n = 0 | IL_0 > 0) \cdot \mathbf{P}(IL_0 > 0).$$

The probability $\mathbf{P}(IL_0 > 0)$ may be estimated from the birth-death process to compute W_0 in the last iteration of the iterative algorithm; we estimate $\mathbf{P}(IL_0 > 0)$ by $\sum_{x=1}^{S_0} \pi_x =: \beta_0$. For the conditional probability $\mathbf{P}(IL_n = 0 | IL_0 > 0)$, we assume that the central warehouse has a strictly positive inventory level for a very long time. Then the behavior of the inventory level at local warehouse n will conform with an Erlang loss system with mean service times t_n instead of LT_n (see the step to compute the fill rates β_n in the iterative algorithm). This leads to:

$$\mathbf{P}(IL_n = 0 | IL_0 > 0) \approx L(S_n, m_n t_n),$$

where $L(\cdot)$ is given by (5). For θ_n, we thus obtain

$$\theta_n \approx \beta_0 L(S_n, m_n t_n). \tag{11}$$

Finally, γ_n can be calculated by substituting β_n and θ_n into (1).

3.2 The method of Muckstadt and Thomas (1980)

We briefly summarize the approximate evaluation method of Muckstadt and Thomas (1980), which is a sequential solution procedure without iterations. It first approximates the mean delay at the central warehouse, and subsequently β_n, θ_n, and γ_n at the local warehouses are computed.

First consider the central warehouse in isolation. They ignore the effect of demands that are fulfilled by the repair facility via an emergency shipment. They assume that the total demand stream is a Poisson process with rate m_0 and the number of backordered demands can grow to infinity. The steady-state behavior is then equal to that of an $M/G/\infty$ queue (see Tijms 2003) with arrival rate m_0 and mean service time t_0. By Palm's theorem, the steady-state probability for x occupied servers within this queueing system equals

$$\pi_x = \frac{(m_0 t_0)^x}{x!} e^{-m_0 t_0}, \quad x \geq 0.$$

The probability distribution of the inventory level IL_0 at the central warehouse is then approximated by:

$$\mathbf{P}(IL_0 = y) = \pi_{S_0 - y}, \quad y \leq S_0.$$

Next, the mean on-hand stock I_0, the mean number of backorders B_0, and the mean delay W_0 are obtained by

$$I_0 = \sum_{y=1}^{S_0} y \pi_{S_0 - y},$$

$$B_0 = \sum_{y=-\infty}^{-1} -y \pi_{S_0 - y} = I_0 - E(IL_0) = I_0 - (S_0 - m_0 t_0),$$

$$W_0 = \frac{B_0}{m_0}.$$

The second step for the computation of the β_n, θ_n, and γ_n proceeds as follows. Per local warehouse $n \in \mathcal{N}$, first the realized replenishment leadtime is approximated by $LT_n = t_n + W_0$ (as in (4) in our method). Then, the fill rate β_n is approximated by $\beta_n = 1 - L(S_n, m_n \cdot LT_n)$ (as in (6)). The fractions θ_n and γ_n are approximated by

$$\theta_n = \beta_0 (1 - \beta_n), \tag{12}$$

$$\gamma_n = (1 - \beta_0)(1 - \beta_n) = 1 - \beta_n - \theta_n, \tag{13}$$

where $\beta_0 = \mathbf{P}(IL_0 > 0) = \sum_{y=1}^{S_0} \pi_{S_0 - y}$.

When comparing our new approximate evaluation method to the method of Muckstadt and Thomas (1980), we see differences at two points:

- The approximation of W_0: There, we use a more refined approximation, where we take into account that unfilled requests for emergency shipments at the central warehouse are satisfied by the repair facility and, thus, they are lost for the central warehouse itself. To incorporate this effect, we make use of the fill rates β_n, and thus we need to iterate.
- The approximation of θ_n: In our approximation, we use (11). Muckstadt and Thomas (1980) use (12), which is equivalent to approximating $\theta_n = \mathbf{P}(IL_0 > 0, IL_n = 0)$ by $\mathbf{P}(IL_0 > 0)\mathbf{P}(IL_n = 0)$, i.e., by assuming that the inventory levels at the central warehouse and local warehouse n behave independently. By using (11), we take a form of dependency into account.

4 Numerical results

This section consists of three parts. We first test the accuracy of our approximation method in Sect. 4.1. After that in Sect. 4.2, we investigate the sensitivity of the system to the repair leadtime distribution. Lastly, we perform a cost optimization experiment in Sect. 4.3.

4.1 Accuracy of the approximate evaluation method

In this subsection, we compare our approximation method with exact results obtained by simulation and with the method of Muckstadt and Thomas (1980).

We consider 96 different instances for our numerical experiment. The input parameters are the number of local warehouses N, the demand rates m_n, the repair lead time t_0, the planned replenishment leadtimes of the local warehouses t_n, and the base stock levels S_n at the central warehouse and all local warehouses. Among all instances, 64 of them are symmetric, where the m_n, S_n, and t_n are the same for all local warehouses, and the remaining 32 instances are asymmetric.

In both the symmetric and asymmetric instances, we consider the following numbers for N: 2, 4, 10, 20. We choose a wide range for N, because there are companies keeping spare parts on stock in only a couple of local warehouses as well as companies with many local warehouses. In the symmetric instances, three different values for m_n are used: 0.01, 0.04, and 0.1 demands per day. We assume $t_n = 3$ days in each instance, and two values are assumed for t_0: 5 and 20 days. The base stock levels S_n are chosen such that the performance measures of the system are within different ranges. We set ϵ, as used for the stopping criteria in our approximation method, at 10^{-6}.

In the asymmetric cases, we determined m_n and t_n for all instances by

$$m_n = m_{n-1} + \Delta_m, \quad n \geq 2,$$
$$t_n = t_{n-1} + \Delta_t, \quad n \geq 2,$$

where Δ_m and Δ_t are chosen constants per instance. The parameters m_1, t_1, Δ_m, and Δ_t of each instance are depicted in Table 1. We choose the base stock levels S_n from the set $\{1, 2, 3\}$, such that they are nondecreasing in n (because m_n is increasing in n). The columns "$S_n = 1$", "$S_n = 2$" and "$S_n = 3$" of Table 1 show the local warehouses with base stock level 1, 2, and 3, respectively. For instance, the value "$1 - 3$" of instance 21 at the column "$S_n = 1$" means that the base stock level is 1 for the local warehouses 1, 2, and 3. The test bed for the asymmetric instances can be seen in Table 1.

We implemented the simulation in the Arena Simulation Software. At each instance, we determined the warm-up period and total run time such that each local warehouse sees at

Table 1 Test bed of the asymmetric instances

Instance	N	m_1	Δ_m	(t_0, t_1)	Δ_t	S_0	$S_n = 1$	$S_n = 2$	$S_n = 3$
1	2	0.01	0.01	(5, 3)	0	1	1–1	2–2	–
2		0.04	0.04	(5, 3)	0	2	1–1	2–2	–
3			0.04	(20, 3)	0	2	1–1	2–2	–
4			0	(5, 2)	1	3	1–1	2–2	–
5		0.1	0.1	(5, 3)	0	3	1–1	2–2	–
6			0.1	(5, 3)	0	4	1–1	2–2	–
7			0.1	(20, 3)	0	4	1–1	2–2	–
8			0	(20, 2)	1	4	1–1	2–2	–
9	4	0.01	0.01	(5, 3)	0	3	1–4	–	–
10		0.04	0.01	(5, 3)	0	3	1–2	3–4	–
11			0.01	(20, 3)	0	3	1–2	3–4	–
12			0	(5, 2)	0.5	2	1–2	3–4	–
13		0.1	0.02	(5, 3)	0	2	1–1	2–4	–
14			0.02	(5, 3)	0	4	1–1	2–4	–
15			0.02	(20, 3)	0	10	1–1	2–3	4–4
16			0	(20, 2)	0.5	5	1–1	2–3	4–4
17	10	0.01	0.01	(5, 3)	0	6	1–8	9–10	–
18		0.04	0.01	(5, 3)	0	10	1–6	7–10	–
19			0.01	(20, 3)	0	20	1–4	5–10	–
20			0	(5, 2)	0.2	4	1–4	5–10	–
21		0.1	0.01	(5, 3)	0	8	1–3	4–10	–
22			0.01	(5, 3)	0	10	1–1	2–8	9–10
23			0.01	(20, 3)	0	25	1–1	2–6	7–10
24			0	(20, 2)	0.2	20	1–1	2–4	5–10
25	20	0.01	0.005	(5, 3)	0	8	1–16	17–20	–
26		0.04	0.005	(5, 3)	0	12	1–12	13–20	–
27			0.005	(20, 3)	0	45	1–10	11–20	–
28			0	(5, 2)	0.1	5	1–15	16–20	–
29		0.1	0.005	(5, 3)	0	20	1–6	7–20	–
30			0.005	(5, 3)	0	15	1–6	7–16	17–20
31			0.005	(20, 3)	0	50	1–6	7–13	14–20
32			0	(20, 2)	0.1	37	1–6	7–15	16–20

least 10,000 demands in the warm-up period and 50,000 demands in total. We performed 100 replications for each instance.

Methods M1, M2, and M3 represent the exact results (via simulation), the results of our approximation method, and the results of the approximation method of Muckstadt and Thomas (1980), respectively. The computation times for the methods M2 and M3 are quite short. All computations have been executed on a computer with an Intel Core2 Duo 2.5 GHz processor. Table 2 shows the average computation times (in milliseconds) for both methods for groups of instances with 2, 4, 10, and 20 local warehouses. We performed 10,000 replications for each instance to get accurate measures for the average computation times. We have not exploited the symmetry when executing the computations in the symmetric instances in

Table 2 Computation times and numbers of iterations

Instances	Average Computation Time (ms)				Average Number of Iterations	
	M2		M3		M2	
	Sym. Inst.	Asym. Inst.	Sym. Inst.	Asym. Inst.	Sym. Inst.	Asym. Inst.
$N = 2$	0.31	0.30	0.03	0.04	4.6	4.4
$N = 4$	0.52	0.56	0.06	0.06	5.6	5.8
$N = 10$	1.12	1.02	0.14	0.16	6.8	5.3
$N = 20$	2.05	1.96	0.27	0.31	7.1	6.0
All N	1.00	0.96	0.12	0.14	6.0	5.3

both methods, i.e., we computed the performance measures for each local warehouse separately in the symmetric instances. According to the results, the average computation times for the methods M2 and M3 are less than 1.00 and 0.14 milliseconds, respectively. We see that the average computation time increases with N for both methods. Table 2 also shows the average number of iterations in the method M2 and we see that the average number of iterations generally increases with N. Further, we see low average number of iterations in both the symmetric and asymmetric instances, i.e. the method M2 converges quickly.

Table 3 shows the results of the symmetric instances and the exact results can be seen with their 95 % confidence intervals. As one can see, the simulated results have been determined with high absolute precision. According to the results, our method M2 clearly outperforms method M3 with respect to the approximation of the β_n. When M3 is accurate, M2 is also accurate. M3 has large deviations from the exact values in several cases, especially when the exact β_n is low, but M2 is still quite accurate in those cases. With respect to the approximation of the θ_n and γ_n, the picture is less clear, but it is clearer when we compute the differences for groups of instances.

In Tables 4 and 5, we see differences between M2 and M1, and M3 and M1, for the symmetric and asymmetric instances, respectively. For groups of instances with 2, 4, 10, and 20 local warehouses, we have computed the average of the difference, the average of the absolute difference, and the maximum absolute difference respectively. In the asymmetric instances, we compute the average of the absolute differences of each performance measure in the following way. We first compute the absolute differences for each performance measure at each warehouse and instance. Then we compute the average values of the absolute differences for all warehouses at each instance, and then take the average over all instances with 2, 4, 10, 20 warehouses.

According to the results, our approximation method M2 is accurate at each of the performance measures. The absolute differences over all instances for β_n, θ_n, and γ_n are less than 0.0067, 0.0129, and 0.0114, respectively. For β_n, the absolute differences are low for all values of N. For θ_n and γ_n, very low absolute differences are obtained for high values of N and larger absolute differences are obtained for low values of N. The latter is most likely due to the stronger dependence between inventory levels at the central and local warehouse(s) when N is low.

With respect to the average difference, our method M2 gives better results than method M3 for β_n and θ_n. However, for γ_n, method M3 gives slightly better results. With respect to average absolute and maximum absolute differences, M2 gives much better results than M3 for all performance measures. This means that our method M2 dominates method M3.

An interesting result is that both M2 and M3 have a tendency to overestimate θ_n. For the symmetric instances, M2 overestimated θ_n at all instances and M3 overestimated θ_n in

Table 3 Results of the symmetric instances

Ins.	N	m_n	(t_0, t_n)	S_0	S_n	β_n			θ_n			γ_n		
						M1	M2	M3	M1	M2	M3	M1	M2	M3
1	2	0.01	(5, 3)	1	1	0.9696 ± 0.0002	0.9686	0.9686	0.0004 ± 0.0000	0.0264	0.0284	0.0300 ± 0.0002	0.0050	0.0030
2			(20, 3)	1	1	0.9454 ± 0.0002	0.9401	0.9388	0.0003 ± 0.0000	0.0196	0.0410	0.0542 ± 0.0002	0.0403	0.0202
3		0.04	(5, 3)	1	1	0.8775 ± 0.0003	0.8671	0.8657	0.0046 ± 0.0001	0.0725	0.0900	0.1179 ± 0.0003	0.0604	0.0443
4				2	1	0.8915 ± 0.0003	0.8895	0.8894	0.0813 ± 0.0002	0.1006	0.1038	0.0272 ± 0.0002	0.0098	0.0068
5				1	2	0.9893 ± 0.0001	0.9897	0.9897	0.0002 ± 0.0000	0.0043	0.0069	0.0105 ± 0.0001	0.0060	0.0034
6			(20, 3)	1	1	0.7164 ± 0.0004	0.7083	0.6575	0.0022 ± 0.0000	0.0278	0.0692	0.2814 ± 0.0004	0.2639	0.2734
7				3	1	0.8704 ± 0.0003	0.8595	0.8510	0.0640 ± 0.0002	0.0855	0.1167	0.0656 ± 0.0002	0.0550	0.0323
8				2	2	0.9645 ± 0.0002	0.9707	0.9697	0.0001 ± 0.0000	0.0034	0.0159	0.0354 ± 0.0002	0.0259	0.0144
9		0.1	(5, 3)	1	1	0.7141 ± 0.0004	0.6894	0.6739	0.0152 ± 0.0001	0.0952	0.1200	0.2707 ± 0.0004	0.2154	0.2061
10				2	1	0.7581 ± 0.0004	0.7445	0.7397	0.1211 ± 0.0003	0.1744	0.1915	0.1209 ± 0.0003	0.0812	0.0688
11				1	2	0.9293 ± 0.0003	0.9282	0.9269	0.0015 ± 0.0000	0.0126	0.0269	0.0692 ± 0.0002	0.0592	0.0462
12				4	2	0.9663 ± 0.0002	0.9661	0.9661	0.0284 ± 0.0002	0.0328	0.0332	0.0053 ± 0.0001	0.0010	0.0006
13			(20, 3)	1	1	0.4428 ± 0.0004	0.4741	0.3560	0.0030 ± 0.0001	0.0206	0.0118	0.5542 ± 0.0004	0.5053	0.6322
14				2	1	0.5592 ± 0.0004	0.5535	0.4246	0.0174 ± 0.0001	0.0520	0.0527	0.4234 ± 0.0004	0.3945	0.5227
15				4	2	0.8964 ± 0.0003	0.8957	0.8764	0.0054 ± 0.0001	0.0159	0.0536	0.0982 ± 0.0003	0.0884	0.0701
16				4	3	0.9629 ± 0.0002	0.9741	0.9723	0.0002 ± 0.0000	0.0015	0.0120	0.0369 ± 0.0002	0.0244	0.0157
17	4	0.01	(5, 3)	1	1	0.9675 ± 0.0002	0.9665	0.9665	0.0011 ± 0.0000	0.0239	0.0274	0.0314 ± 0.0002	0.0096	0.0061
18			(20, 3)	1	1	0.9229 ± 0.0003	0.9170	0.9155	0.0007 ± 0.0000	0.0134	0.0380	0.0765 ± 0.0003	0.0697	0.0465
19		0.04	(5, 3)	1	1	0.8567 ± 0.0003	0.8480	0.8458	0.0087 ± 0.0001	0.0500	0.0693	0.1346 ± 0.0003	0.1020	0.0849
20				3	1	0.8918 ± 0.0003	0.8908	0.8907	0.0919 ± 0.0003	0.1022	0.1041	0.0163 ± 0.0001	0.0070	0.0052
21				2	2	0.9920 ± 0.0001	0.9921	0.9921	0.0003 ± 0.0000	0.0052	0.0064	0.0077 ± 0.0001	0.0027	0.0015
22			(20, 3)	1	1	0.6409 ± 0.0004	0.6369	0.5952	0.0016 ± 0.0000	0.0095	0.0165	0.3575 ± 0.0004	0.3536	0.3883
23				3	1	0.8082 ± 0.0004	0.7853	0.7587	0.0277 ± 0.0002	0.0477	0.0917	0.1641 ± 0.0004	0.1670	0.1496
24				3	2	0.9591 ± 0.0002	0.9644	0.9631	0.0005 ± 0.0000	0.0025	0.0140	0.0404 ± 0.0002	0.0332	0.0229
25		0.1	(5, 3)	1	1	0.6635 ± 0.0004	0.6498	0.6314	0.0167 ± 0.0001	0.0453	0.0499	0.3198 ± 0.0004	0.3049	0.3187
26				2	1	0.7272 ± 0.0004	0.7094	0.6967	0.0634 ± 0.0002	0.1082	0.1231	0.2094 ± 0.0004	0.1825	0.1802
27				4	2	0.9636 ± 0.0002	0.9630	0.9629	0.0179 ± 0.0001	0.0288	0.0318	0.0184 ± 0.0001	0.0082	0.0053
28				4	3	0.9959 ± 0.0001	0.9961	0.9961	0.0008 ± 0.0000	0.0029	0.0034	0.0033 ± 0.0001	0.0011	0.0006
29			(20, 3)	1	1	0.3778 ± 0.0003	0.4033	0.3279	0.0008 ± 0.0000	0.0043	0.0002	0.6214 ± 0.0003	0.5925	0.6719
30				2	1	0.4492 ± 0.0003	0.4597	0.3570	0.0039 ± 0.0001	0.0131	0.0019	0.5468 ± 0.0004	0.5272	0.6410
31				4	2	0.7790 ± 0.0004	0.7798	0.7281	0.0008 ± 0.0000	0.0032	0.0115	0.2201 ± 0.0004	0.2170	0.2604
32				6	3	0.9468 ± 0.0003	0.9557	0.9514	0.0002 ± 0.0000	0.0007	0.0093	0.0530 ± 0.0003	0.0436	0.0393

Table 3 (*Continued*)

Ins.	N	m_n	(t_0, t_n)	S_0	S_n	β_n			θ_n			γ_n		
						M1	M2	M3	M1	M2	M3	M1	M2	M3
33	10	0.01	(5, 3)	2	1	0.9696 ± 0.0002	0.9693	0.9693	0.0189 ± 0.0001	0.0265	0.0279	0.0115 ± 0.0001	0.0041	0.0028
34			(20, 3)	3	1	0.9540 ± 0.0002	0.9515	0.9507	0.0131 ± 0.0001	0.0199	0.0333	0.0329 ± 0.0002	0.0286	0.0159
35		0.04	(5, 3)	1	1	0.8200 ± 0.0004	0.8162	0.8107	0.0071 ± 0.0001	0.0177	0.0256	0.1729 ± 0.0004	0.1661	0.1637
36				3	1	0.8807 ± 0.0003	0.8773	0.8758	0.0545 ± 0.0002	0.0744	0.0840	0.0648 ± 0.0002	0.0483	0.0402
37				4	2	0.9928 ± 0.0001	0.9928	0.9928	0.0031 ± 0.0001	0.0055	0.0061	0.0041 ± 0.0001	0.0017	0.0010
38			(20, 3)	4	1	0.7191 ± 0.0004	0.7062	0.6553	0.0063 ± 0.0001	0.0128	0.0146	0.2746 ± 0.0004	0.2810	0.3301
39				10	1	0.8751 ± 0.0003	0.8671	0.8602	0.0733 ± 0.0003	0.0807	0.1002	0.0515 ± 0.0002	0.0521	0.0396
40				6	2	0.9545 ± 0.0002	0.9584	0.9556	0.0005 ± 0.0000	0.0013	0.0085	0.0450 ± 0.0002	0.0403	0.0359
41		0.1	(5, 3)	2	1	0.6563 ± 0.0004	0.6496	0.6232	0.0150 ± 0.0001	0.0243	0.0152	0.3287 ± 0.0004	0.3261	0.3616
42				5	1	0.7440 ± 0.0004	0.7345	0.7206	0.1019 ± 0.0003	0.1237	0.1231	0.1541 ± 0.0004	0.1418	0.1563
43				8	2	0.9646 ± 0.0002	0.9643	0.9642	0.0237 ± 0.0002	0.0291	0.0310	0.0117 ± 0.0001	0.0066	0.0048
44				10	3	0.9966 ± 0.0001	0.9966	0.9966	0.0027 ± 0.0001	0.0032	0.0033	0.0007 ± 0.0000	0.0002	0.0001
45			(20, 3)	10	1	0.5911 ± 0.0004	0.5705	0.4346	0.0169 ± 0.0001	0.0285	0.0028	0.3921 ± 0.0004	0.4010	0.5625
46				16	1	0.7047 ± 0.0004	0.6774	0.5745	0.0828 ± 0.0003	0.0936	0.0666	0.2125 ± 0.0004	0.2290	0.3589
47				17	2	0.9079 ± 0.0003	0.9043	0.8833	0.0068 ± 0.0001	0.0100	0.0258	0.0853 ± 0.0003	0.0856	0.0909
48				25	2	0.9613 ± 0.0002	0.9608	0.9601	0.0263 ± 0.0002	0.0286	0.0337	0.0124 ± 0.0001	0.0106	0.0063
49	20	0.01	(5, 3)	3	1	0.9699 ± 0.0002	0.9698	0.9698	0.0224 ± 0.0001	0.0268	0.0278	0.0077 ± 0.0001	0.0034	0.0024
50			(20, 3)	3	1	0.9186 ± 0.0003	0.9155	0.9112	0.0040 ± 0.0001	0.0077	0.0211	0.0774 ± 0.0003	0.0767	0.0676
51		0.04	(5, 3)	2	1	0.8263 ± 0.0003	0.8240	0.8160	0.0085 ± 0.0001	0.0139	0.0169	0.1653 ± 0.0003	0.1621	0.1672
52				5	1	0.8813 ± 0.0003	0.8789	0.8768	0.0578 ± 0.0002	0.0706	0.0775	0.0609 ± 0.0002	0.0505	0.0457
53				7	2	0.9931 ± 0.0001	0.9932	0.9932	0.0044 ± 0.0001	0.0057	0.0061	0.0025 ± 0.0001	0.0011	0.0008
54			(20, 3)	10	1	0.7708 ± 0.0004	0.7598	0.7022	0.0101 ± 0.0001	0.0154	0.0129	0.2191 ± 0.0004	0.2248	0.2849
55				20	1	0.8854 ± 0.0003	0.8823	0.8784	0.0871 ± 0.0003	0.0903	0.0987	0.0275 ± 0.0002	0.0274	0.0228
56				15	2	0.9783 ± 0.0002	0.9804	0.9796	0.0016 ± 0.0000	0.0024	0.0075	0.0201 ± 0.0001	0.0171	0.0129
57		0.1	(5, 3)	5	1	0.6847 ± 0.0004	0.6796	0.6443	0.0195 ± 0.0001	0.0264	0.0104	0.2958 ± 0.0004	0.2940	0.3453
58				10	1	0.7534 ± 0.0004	0.7480	0.7339	0.1227 ± 0.0003	0.1350	0.1218	0.1239 ± 0.0003	0.1170	0.1442
59				12	2	0.9620 ± 0.0002	0.9618	0.9614	0.0191 ± 0.0001	0.0238	0.0269	0.0189 ± 0.0001	0.0144	0.0117
60				15	3	0.9965 ± 0.0001	0.9965	0.9965	0.0025 ± 0.0001	0.0031	0.0032	0.0010 ± 0.0000	0.0004	0.0003
61			(20, 3)	30	1	0.7053 ± 0.0004	0.6850	0.5537	0.0641 ± 0.0002	0.0718	0.0193	0.2305 ± 0.0004	0.2433	0.4270
62				40	1	0.7544 ± 0.0004	0.7457	0.7013	0.1596 ± 0.0003	0.1613	0.1431	0.0860 ± 0.0003	0.0930	0.1556
63				30	2	0.8869 ± 0.0003	0.8855	0.8475	0.0024 ± 0.0001	0.0039	0.0066	0.1107 ± 0.0003	0.1107	0.1459
64				40	2	0.9481 ± 0.0002	0.9464	0.9402	0.0160 ± 0.0001	0.0181	0.0286	0.0360 ± 0.0002	0.0355	0.0312

Table 4 Results of the symmetric instances for groups of instances

Instances	β_n		θ_n		γ_n	
	M2-M1	M3-M1	M2-M1	M3-M1	M2-M1	M3-M1
	Average Difference					
$N=2$	−0.0022	−0.0242	0.0250	0.0393	−0.0228	−0.0151
$N=4$	−0.0015	−0.0227	0.0140	0.0226	−0.0124	0.0001
$N=10$	−0.0060	−0.0291	0.0079	0.0093	−0.0020	0.0197
$N=20$	−0.0039	−0.0256	0.0046	0.0017	−0.0007	0.0239
All N	−0.0034	−0.0254	0.0129	0.0182	−0.0095	0.0072
	Average Absolute Difference					
$N=2$	0.0083	0.0261	0.0250	0.0393	0.0228	0.0372
$N=4$	0.0079	0.0238	0.0140	0.0229	0.0128	0.0269
$N=10$	0.0065	0.0292	0.0079	0.0131	0.0061	0.0319
$N=20$	0.0042	0.0257	0.0046	0.0106	0.0039	0.0310
All N	0.0067	0.0262	0.0129	0.0215	0.0114	0.0317
	Maximum Absolute Difference					
$N=2$	0.0313	0.1345	0.0800	0.1048	0.0575	0.0993
$N=4$	0.0255	0.0922	0.0448	0.0640	0.0326	0.0942
$N=10$	0.0274	0.1565	0.0218	0.0295	0.0166	0.1705
$N=20$	0.0204	0.1517	0.0128	0.0448	0.0127	0.1965
All N	0.0313	0.1565	0.0800	0.1048	0.0575	0.1965

56 of the 64 instances. Similar results can be observed for the asymmetric instances (we see this result when looking at the underlying values for the θ_n). We can analyze this result by considering (11) and (12). In Table 6, the average difference, the average of the absolute difference, and the maximum absolute difference between M2 and M1, and M3 and M1 with respect to β_0 are given.

Firstly, we see from Table 6 that M2 and M3 underestimate β_0 on average. (We can also see this result in the underlying β_0 values, i.e., M2 underestimates β_0 at all instances except the symmetric instance 63, and M3 underestimates β_0 at all 96 instances. For more detailed results see Özkan et al. (2011), which is the working paper version of this study and contains more detailed numerical results.) We explain this result in the following way. The approximation method M2 is based on the implicit assumption that the inventory levels at the local warehouses are independent of the inventory level at the central warehouse. More precisely, when analyzing the behavior of the central warehouse, we assume that, independent of the actual inventory level, there is always a Poisson demand stream with rate $m_n\beta_n$ for replenishment orders placed by local warehouse n and a Poisson demand stream with rate $m_n(1-\beta_n)$ for emergency shipment requests placed by local warehouse n ($n \in \mathcal{N}$). This leads to the approximate birth-death process for the behavior of the inventory level at the central warehouse as depicted in Fig. 2. However, in the true system, we have a positive correlation between the inventory level IL_0 at the central warehouse and the inventory levels IL_n at the local warehouses $n \in \mathcal{N}$. Hence, in the true system, the total stream of emergency shipment requests will have a higher rate than $\sum_{n\in\mathcal{N}} m_n(1-\beta_n)$ when $IL_0 \leq 0$, and the stream of replenishment orders will have a lower rate than $\sum_{n\in\mathcal{N}} m_n\beta_n$. Hence, in Fig. 2, the rate m_0' for transitions to the left when $IL_0 \leq 0$ is an overestimation, and this leads to an

Table 5 Results of the asymmetric instances for groups of instances

Instances	β_n		θ_n		γ_n	
	M2-M1	M3-M1	M2-M1	M3-M1	M2-M1	M3-M1
	Average Difference					
$N=2$	−0.0065	−0.0316	0.0189	0.0364	−0.0123	−0.0047
$N=4$	−0.0054	−0.0182	0.0113	0.0239	−0.0059	−0.0057
$N=10$	−0.0019	−0.0072	0.0053	0.0116	−0.0034	−0.0044
$N=20$	−0.0025	−0.0131	0.0052	0.0073	−0.0026	0.0058
All N	−0.0041	−0.0175	0.0102	0.0198	−0.0061	−0.0023
	Average Absolute Difference					
$N=2$	0.0075	0.0317	0.0189	0.0364	0.0141	0.0268
$N=4$	0.0065	0.0182	0.0113	0.0239	0.0097	0.0157
$N=10$	0.0024	0.0073	0.0053	0.0116	0.0048	0.0089
$N=20$	0.0026	0.0131	0.0052	0.0089	0.0049	0.0147
All N	0.0048	0.0176	0.0102	0.0202	0.0084	0.0165
	Maximum Absolute Difference					
$N=2$	0.0425	0.1431	0.0381	0.1034	0.0308	0.0884
$N=4$	0.0380	0.0941	0.0292	0.0524	0.0252	0.0615
$N=10$	0.0233	0.0625	0.0167	0.0510	0.0163	0.0483
$N=20$	0.0191	0.0870	0.0142	0.0238	0.0152	0.1108
All N	0.0425	0.1431	0.0381	0.1034	0.0308	0.1108

Table 6 Average, average of the absolute, and maximum absolute differences for β_0

	Symmetric Ins.		Asymmetric Ins.	
	M2-M1	M3-M1	M2-M1	M3-M1
Average Diff.	−0.0435	−0.0864	−0.0394	−0.0791
Absolute Diff.	0.0437	0.0864	0.0394	0.0791
Maximum Abs. Diff.	0.1559	0.3865	0.1105	0.2462

underestimation of β_0. Notice that the bounding of the state space (at state $-\bar{S}$) reduces the effect of the overestimation of the transitions to the left when $IL_0 \leq 0$.

Method M3 underestimates β_0 more than M2, as seen in Table 6. This is explained by the fact that M3 assumes that the demand rate at the central warehouse is always equal to m_0. Hence, the transitions to the left when $IL_0 \leq 0$ are even further overestimated. Furthermore, no bounding of the state space is assumed.

Although method M2 generally underestimates β_0, it overestimates θ_n, which is determined via (11). The reason is that $L(S_n, m_n t_n)$ generally overestimates $\mathbf{P}(IL_n = 0 \mid IL_0 > 0)$. In the true system, there is a positive correlation between IL_0 and IL_n. If there is positive stock at the central warehouse, then it is less likely to have zero stock at a local warehouse. Apparently, the relative overestimation of $\mathbf{P}(IL_n = 0 | IL_0 > 0)$ is larger than the relative underestimation of β_0. Lastly, because method M2 generally overestimates θ_n, it has a tendency to underestimate γ_n because of (1). Similarly, although M3 underestimates β_0 in all instances, it generally overestimates θ_n, which is determined by (12). The reason is that

Table 7 Average, average absolute, and maximum absolute differences between the deterministic and the remaining distribution cases

Case	Average Difference			Average Absolute Diff.			Maximum Absolute Diff.		
	β_n	θ_n	γ_n	β_n	θ_n	γ_n	β_n	θ_n	γ_n
Erlang-Det.	−0.0004	0.0006	−0.0002	0.0004	0.0006	0.0004	0.0024	0.0044	0.0020
Expo.-Det.	−0.0008	0.0022	−0.0014	0.0008	0.0022	0.0015	0.0048	0.0125	0.0081
Log.-Det.	−0.0010	0.0028	−0.0018	0.0010	0.0028	0.0020	0.0060	0.0161	0.0110

$1 - \beta_n$ severely overestimates $\mathbf{P}(IL_n = 0 | IL_0 > 0)$. Again, this is because of the ignored positive correlation between IL_0 and IL_n.

Notice that all over- and underestimations become smaller when the correlation between IL_0 and the IL_n is not as pronounced. This is typically so when we have higher numbers of local warehouses.

4.2 Sensitivity analysis

In our model, we assumed that the repair leadtime at the repair facility is deterministic. This leadtime has been denoted by t_0. However, this assumption does not always hold in a real-life situation as there may be variability in the leadtime. Here, we analyze the sensitivity of the system performance with respect to the repair leadtime distribution. We consider four different distributions. The first distribution is the *deterministic distribution*, cf. the assumption in our model. As we mentioned in Sect. 3, we assumed an exponential distribution for the repair leadtime in our approximate evaluation method (in the step to determine the mean delay W_0). So, the second distribution that we consider is the *exponential distribution*, with mean time t_0. The other two distributions are with a coefficient of variation of 0.5 and 2, respectively. We choose an *Erlang-4 distribution* as the third distribution. This distribution has a coefficient of variation of 0.5; its scale parameter is chosen such that the mean is equal to t_0. For the fourth distribution, we choose a *lognormal distribution*, with parameters such that the coefficient of variation is 2 and the mean equals t_0. We simulated results for each performance measure under each distribution, and generated results for the symmetric instances with 4 and 10 local warehouses (32 instances in total). Table 7 depicts the average difference, average absolute difference, and maximum absolute difference of the deterministic case with the other distributions (for more detailed results, see Özkan et al. 2011).

According to the Table 7, the average differences and average absolute differences are all below 0.003. Hence, we may conclude that the performance is rather insensitive to the repair leadtime distribution. We also made some simulation runs for the asymmetric instances to check the sensitivity, and we got similar results. This implies that our approximate evaluation method works also well for systems with a generally distributed repair leadtime.

Another interesting result that we see from Table 7 is that the most sensitive performance measure for the repair leadtime is θ_n, which is also the one that our approximation method estimates the worst among the three performance measures β_n, θ_n, and γ_n. Table 7 also shows that the insensitivity of the system to the repair leadtime decreases as the coefficient of the variation increases. This means that our system is not completely insensitive to the repair leadtime distribution, but even in the lognormal distribution case, which has the highest coefficient of variation among the four distribution cases; the average, average absolute, and maximum absolute differences are still very low.

4.3 Cost optimization

In this section, we demonstrate the use of our approximation method for a single-item optimization problem. The objective is to minimize the total cost subject to mean waiting time constraint per local warehouse. Let W_n^{obj} denote the target mean waiting time at local warehouse $n \in \mathcal{N}$. Then, our optimization problem is as follows:

$$\begin{aligned} &\min\ C \\ &\text{s.t.}\quad W_n \le W_n^{obj}, \quad \forall n \in \mathcal{N}, \\ &S_n \in \mathbb{N} \cup \{0\}, \quad \forall n \in \mathcal{N} \cup \{0\}, \end{aligned}$$

where C and W_n are defined as in (3) and (2), respectively.

The above optimization problem may be solved by a smart enumeration method. First, we derive lower bounds for the base stock levels at the local warehouses under a feasible solution. By (2) and (1), $W_n \ge (1-\beta_n)\hat{t}_n$, where $\hat{t}_n := \min\{t_n^{CW}, t_n^{RF}\}$ (generally, it will hold that $t_n^{CW} \le t_n^{RF}$ and then $\hat{t}_n = t_n^{CW}$). By (4) and (6), $LT_n \ge t_n$ and thus $\beta_n \le 1 - L(S_n, m_n t_n)$ (here, we use the property that $L(c,\rho)$ is increasing as a function of ρ, cf. Harel 1990). Hence, $W_n \ge \hat{t}_n L(S_n, m_n t_n)$, and each feasible solution $(S_0, S_1, \ldots, S_N)$ satisfies $S_n \ge s_n$, $n \in \mathcal{N}$, where:

$$s_n = \min\bigl\{ j \in \mathbb{N} \cup \{0\} : \hat{t}_n L(j, m_n t_n) \le W_n^{obj} \bigr\} \tag{14}$$

(here, we use that $L(c,\rho)$ is decreasing as a function of $c \ge 0$, cf. Karush 1957; see also Remark 2 in Kranenburg and van Houtum 2007).

Next, define $C(k)$ as the solution with the lowest costs of all feasible solutions with a total stock of $k = \sum_{n=0}^{N} S_n$ units. Because of the above lower bounds s_n for the base stock levels at the local warehouses, there is no feasible solution for $k < \sum_{n=1}^{N} s_n$. An optimal solution may be computed by determining $C(k)$ for $k = \sum_{n=1}^{N} s_n, \sum_{n=1}^{N} s_n + 1, \ldots$. Obviously, $C(k) \ge hk$ for all k. Hence, this procedure may be stopped at a given value for k as soon as the cost of the best solution under a total stock of at most $k = \sum_{n=0}^{N} S_n$ units, is less than or equal to $h(k+1)$, i.e., as soon as $C^*(k) \le h(k+1)$ with $C^*(k) := \min_{j \le k} C(j)$. This leads to the following exact solution procedure:

Step 0 Let $k = \sum_{n=1}^{N} s_n$, $C^*(k) = \infty$.
Step 1 For each $(S_0, S_1, \ldots, S_N)$ with $\sum_{n=0}^{N} S_n = k$ and $S_n \ge s_n$ for all $n \in \mathcal{N}$, compute C and W_n for all $n \in \mathcal{N}$. If $C < C^*(k)$ and $W_n \le W_n^{obj}$ for all $n \in \mathcal{N}$, then $C^*(k) = C$.
Step 2 If $C^*(k) \le h(k+1)$, then stop, else $k := k+1$ and go to *Step 1*.

The above procedure generates an optimal solution when an exact evaluation method is used. We use the above method with the approximate evaluation method M2, which leads to a heuristic solution. Because of the accuracy of method M2, one may expect that the heuristic solutions will be close to optimal and that their mean waiting times are below or close to the target levels. We test this in a small experiment consisting of 10 asymmetric instances with $N = 6$ local warehouses. We apply the above smart enumeration method with evaluations by method M2. This leads to a heuristic solution $(S_0, S_1, \ldots, S_N)$, approximated costs C and approximated mean waiting times W_n, $n \in \mathcal{N}$. Next by simulation, we determine the exact costs and waiting times of the heuristic solution. We compare the exact costs to the approximated costs and we measure

$$\Delta_W = \sum_{n=1}^{N} \max\bigl\{0, W_n - W_n^{obj}\bigr\};$$

Table 8 Results of the cost optimization experiment

Inst.	N	m_1	Δ_m	(t_0, t_1)	Δ_t	h	Heuristic sol. $(S_0, S_1, \ldots, S_6)$	Total Costs (C)			Feasib. Δ_W (hours)
								M1	M2	M2-M1	
1	6	0.01	0.01	(5, 3)	0	20	(2, 1, 1, 1, 2, 2, 2)	226	225	0.4 %	0.013
2		0.04	0.01	(5, 3)	0	20	(2, 2, 2, 2, 2, 2, 2)	293	292	0.5 %	0
3			0.01	(20, 3)	0	2	(9, 2, 2, 2, 2, 3, 3)	53.5	52.1	2.7 %	0
4			0	(5, 2)	0.5	2	(2, 2, 2, 2, 2, 2, 2)	30.6	30.0	2.1 %	0
5			0.01	(20, 2)	0.5	20	(8, 2, 2, 2, 2, 2, 3)	436	435	0.3 %	0
6		0.08	0.01	(5, 3)	0	20	(4, 2, 2, 2, 2, 2, 2)	348	345	0.9 %	0
7			0.01	(20, 3)	0	2	(14, 3, 3, 3, 3, 3, 3)	72.5	70.0	3.4 %	0
8			0	(20, 2)	0.5	2	(10, 3, 3, 3, 3, 3, 3)	62.1	60.0	3.3 %	0
9			0.01	(5, 2)	0.5	2	(5, 2, 2, 3, 3, 3, 3)	49.5	47.9	3.2 %	0
10			0.01	(20, 2)	0.5	20	(13, 2, 2, 2, 3, 3, 3)	584	582	0.3 %	0

Δ_W measures how much the mean waiting time constraints are violated.

For all 10 instances, we take $t_n^{CW} = 10$ hours, $t_n^{RF} = 20$ hours, $C_n^{CW} = 500$ USD, $C_n^{RF} = 1000$ USD, and $W_n^{obj} = 1.5$ hours for all $n \in \mathcal{N}$. For h, we take 2 and 20 USD per unit per day, which corresponds to items with a price of 5,000 and 50,000 USD, respectively. The other parameters are denoted in the same way as for the asymmetric instances in Sect. 4.1. The results are listed in Table 8.

According to the results, total cost values of the method M1 are close to the results of M2. The average absolute deviation of M2 from M1 is 1.7 %. Moreover, M1 gives feasible results at each instance except instance 1. In instance 1, W_3 is just a little bit larger than W_3^{obj} (the difference is less than one minute); this difference can be considered as negligible in practical applications. When $h = 20$, i.e. the inventory holding cost is high, the absolute differences of the method M2 with respect to M1 are smaller than for $h = 2$. This result is expected, because the inventory holding cost of the two methods are the same and the only cost difference occurs in the computation of total emergency shipment costs. When $h = 20$, total inventory holding costs dominate the total emergency shipment costs in each of the methods, and thus the relative absolute differences with respect to the total cost values are smaller. Another interesting observation is that the method M2 underestimates the total cost values at each instance. The reason of this result is that M2 has a tendency to overestimate θ_n and underestimate γ_n as explained in Sect. 4.1, and $C_n^{RF} > C_n^{CW}$ for all $n \in \mathcal{N}$ in our experiment. Because the results of the method M1 are close to the method M2 with respect to the objective function value and the use of M2 in the smart enumeration procedure leads to feasible solutions, we may say that M2 can be safely used for this type of optimization problems.

Note that, for given k, N, and s_n, $n \in \mathcal{N}$, the number of solutions considered in Step 1 of the smart enumeration procedure is equal to $(k - \sum_{n=1}^{N} s_n + N)!/[(k - \sum_{n=1}^{N} s_n)!N!]$. Therefore, as k and N increase, the number of solutions and the computation time grow exponentially. Hence, for large problems, one has to use other procedures such as greedy procedures; see e.g. the greedy procedures in Wong et al. (2005, 2007) and Kranenburg and van Houtum (2009), which have been shown to work well for similar optimization problems.

5 Conclusion

In this study, we derived an accurate and fast approximate evaluation method for two-echelon spare parts systems with emergency shipments. We also showed that our method outperforms the method of Muckstadt and Thomas (1980). Further, we showed that the performance measures of our system are virtually insensitive to the repair leadtime distribution, which increases the applicability of our approximation method. Lastly, we performed a cost optimization experiment and show that our approximation method can be safely used in optimization problems as well.

The main idea behind our approximation method is to decompose the analysis of the whole system into an analysis for local warehouses (leading to the β_n for a given W_0) and an analysis for the central warehouse (leading to W_0 for given β_n's), and an iterative procedure to couple those two analyses. This idea may also work for systems with additional features.

One of such features is that, demands do not only occur for single units at a time but for two or more units ('compounds'). One has this feature when a repairable occurs multiple times in the configuration of a technical system. It may be desired to replace all parts of a repairable when one of them fails. Modeling demand as compound Poisson processes will be more accurate in that case. The main idea of our approximation method can still be followed, but the procedure has to be adapted at multiple places (notice that one also has to specify whether partial or only complete fulfillments of demands are allowed by a local warehouse): (i) in the analysis of a local warehouse, one gets a parallel with an $M^X/G/c/c$ instead of an $M/G/c/c$ queue and one gets more complicated calculations for the 'overflow' demand processes to the central warehouse and the repair facility; (ii) in the analysis of the central warehouse, one does not get a birth-death process anymore, but a Markov process with a more general transition structure, which requires a computational solution; (iii) the logic to derive θ_n has to be adapted. Obviously, new numerical experiments are needed to verify whether the resulting approximation method would still be accurate.

Another additional feature that one may have in a real-life situation is the presence of lateral shipments, where the application of a lateral shipment may be preferred above an emergency shipment from the central warehouse (as in Alfredsson and Verrijdt 1999) or the other way around (because of logistics reasons). The main idea of our method may also work in that case, but in the analysis the 'overflow' demand streams because of the lateral shipments have to be added (e.g., like in Kranenburg and van Houtum 2009).

Acknowledgements The authors would like to thank the guest editor, anonymous referees, and Dr. Jeffrey Kharoufeh for their helpful comments on earlier versions of this paper.

References

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

Alvarez, E., & van der Heijden, M. (2011). *On two-echelon inventory systems with Poisson demand and lost sales*. Beta working paper, 366. Available at http://beta.ieis.tue.nl/publications/workingpapers.

Andersson, J., & Melchiors, P. (2001). A two-echelon inventory model with lost sales. *International Journal of Production Economics*, *69*, 307–315.

Axsäter, S. (1990). Modelling emergency lateral transshipments in inventory systems. *Management Science*, *36*, 1329–1338.

Axsäter, S., Kleijn, M., & De Kok, T. G. (2004). Stock rationing in a continuous review two-echelon inventory model. *Annals of Operations Research*, *126*, 177–194.

Basten, R. J. I. (2010). *Designing logistics support systems: level of repair analysis and spare parts inventories*. PhD thesis, University of Twente.

Basten, R. J. I., Schutten, J. M. J., & van der Heijden, M. C. (2009). An efficient model formulation for level of repair analysis. *Annals of Operations Research*, *172*, 119–142.

Grahovac, J., & Chakravarty, A. (2001). Sharing and lateral transshipments of inventory in a supply chain with expensive low-demand items. *Management Science*, *47*, 579–594.

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

Harel, A. (1990). Convexity properties of Erlang loss formula. *Operations Research*, *38*, 499–505.

Hausman, W. H., & Erkip, N. K. (1994). Multi-echelon vs. single-echelon inventory control policies for low-demand items. *Management Science*, *40*, 597–602.

Karush, W. (1957). A queueing model for an inventory problem. *Operations Research*, *5*, 693–703.

Kranenburg, A. A., & van Houtum, G. J. (2007). Cost optimization in the $(S-1, S)$ lost sales inventory model with multiple demand classes. *Operations Research Letters*, *35*, 493–502.

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

Muckstadt, J. A., & Thomas, L. J. (1980). Are multi-echelon inventory methods worth implementing in systems with low-demand-rate-items? *Management Science*, *26*, 483–494.

Özkan, E., van Houtum, G. J., & Serin, Y. (2011). *A new approximate evaluation method for two-echelon inventory systems with emergency shipments*. Beta working paper, 363. Available at http://beta.ieis.tue.nl/publications/workingpapers.

Paterson, C., Kiesmüller, G., Teunter, R., & Glazebrook, K. (2011). Inventory models with lateral transshipments: a review. *European Journal of Operational Research*, *210*, 125–136.

Rustenburg, W. D., van Houtum, G. J., & Zijm, W. H. M. (2003). Exact and approximate analysis of multi-echelon, multi-indenture spare parts systems with commonality. In J. G. Shanthikumar, D. D. Yao, & W. H. M. Zijm (Eds.), *Stochastic modeling and optimization of manufacturing systems and supply chains* (pp. 143–176). Boston: Kluwer Academic.

Saranga, H., & Kumar, U. D. (2006). Optimization of aircraft maintenance/support infrastructure using genetic algorithms—level of repair analysis. *Annals of Operations Research*, *143*, 91–106.

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

Tijms, H. J. (2003). *A first course in stochastic models* (2nd ed.). New York: Wiley.

Wong, H., van Houtum, G. J., Cattrysse, D., & van Oudheusden, D. (2005). Simple, efficient heuristics for multi-item, multi-location spare parts systems with lateral transshipments and waiting time constraints. *Journal of the Operational Research Society*, *56*, 1419–1430.

Wong, H., Kranenburg, B., van Houtum, G. J., & Cattrysse, D. (2007). Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse. *OR Spektrum*, *29*, 699–722.