University of Groningen

Optimizing stock levels for rental systems with a support warehouse and partial backordering

van der Heide, Gerlach; van Foreest, Nicky; Roodbergen, Kees Jan

Published in:
European Journal of Operational Research

DOI:
[10.1016/j.ejor.2017.07.040](https://doi.org/10.1016/j.ejor.2017.07.040)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

van der Heide, G., van Foreest, N., & Roodbergen, K. J. (2018). Optimizing stock levels for rental systems with a support warehouse and partial backordering. *European Journal of Operational Research*, 265(1), 107-118. <https://doi.org/10.1016/j.ejor.2017.07.040>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Production, Manufacturing and Logistics

Optimizing stock levels for rental systems with a support warehouse and partial backordering

G. Van der Heide*, N. D. Van Foreest, K. J. Roodbergen

University of Groningen, P.O. Box 800, Groningen 9700 AV, The Netherlands

ARTICLE INFO

Article history:

Received 15 July 2016

Accepted 12 July 2017

Available online 19 July 2017

Keywords:

Inventory

Rental system

Support warehouse

Partial backordering

Lateral transshipments

ABSTRACT

Various rental systems, such as public libraries and tool rental companies, have a support warehouse for storing rental products at low cost and carrying out expedited shipments in response to stock-outs at local warehouses. In case of stock-outs, rental systems commonly employ partial backordering, i.e., a limited number of demands is backordered while additional demand is lost. We include partial backordering in our models and demonstrate its relevance. We fully characterize costs and optimal base stock policies in cases with one support and one local warehouse and provide upper bounds for base stock levels in the general system. By enumerating base stock policies and evaluating exact costs numerically, we determine optimal base stock policies for small systems with up to six local warehouses. For larger systems we develop a greedy algorithm based on approximate cost evaluations. The resulting base stock policies have costs within 0.2% from optimality on average in experiments. Among other things, experiments indicate that existing methods with no backordering may lead to poor solutions with partial backordering and that partial backordering is an effective means to reduce lost demand even with limited partial backorder levels.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

In rental systems with multiple local warehouses, rental products are sometimes stored in a support warehouse and shipped on demand. This occurs especially in systems with many different items which each are demanded with low probability at local warehouses. For example, public libraries have one or several copies of a book on shelf, additional copies can be delivered from a central library if needed. A similar situation is faced by tool rental companies that store extra tools in a support warehouse for shipment when required at one of the local warehouses. In both settings, it is relevant to optimize stock levels at the local and support warehouses.

Shipments from a support warehouse are an alternative to lateral transshipment between all local warehouses. Support warehouses provide economies of scale since numerous items are handled in one dedicated place with low storage costs. This requires less coordination and handling than comparable lateral transshipment systems. Furthermore, the use of a support warehouse is preferred over lateral transshipments if shipment costs are high (Rosales, Rao, & Rogers, 2013). For public libraries these are

important considerations since the demand for physical books has decreased significantly in the last two decades. Low demand emphasizes a focus on low-cost storage rather than fast shipments. And as shipment volumes decrease, shipment cost per item are increasing. While traditionally lateral transshipments are employed (Van der Heide & Roodbergen, 2013), a transition is therefore made towards centralized operations in support warehouses to consolidate the progressively smaller demand streams and to increase utilization of infrastructure.

In this paper, we consider a rental system with several local warehouses and a support warehouse dedicated to storage and expedited shipments. Each warehouse specifies its number of rental items by setting a base stock level. The support warehouse has (non-strictly) lower holding costs than the local warehouses. Local warehouses face Poisson demand and serve this demand from local stock. In out-of-stock situations the demand is fulfilled by carrying out expedited shipments from the support warehouse. If the support warehouse is also out of stock, demand is backordered at the local warehouse up to a specified level R and lost otherwise, i.e., there is partial backordering. Partial backordering has no backordering ($R = 0$) and full backordering ($R = \infty$) as special cases.

Due to limited rental stock, partial backordering is a natural feature of rental systems. For example, public library systems typically allow one or two backorders per book title per library and reject any additional book requests. More generally, when rental

* Corresponding author.

E-mail addresses: g.van.der.heide@rug.nl (G. Van der Heide),
n.d.van.foreest@rug.nl (N.D. Van Foreest), k.j.roodbergen@rug.nl (K.J. Roodbergen).

Table 1
Overview of literature about support warehouses.

Article	Demand	Policy	Backorders
Axsäter et al. (2013)	Normal	(R, Q)	Full
Howard et al. (2015)	Poisson	Base stock	Threshold
Kranenburg and Van Houtum (2009)	Poisson	Base stock	None
Van Wijk et al. (2013)	Poisson	Base stock	None
This article	Poisson	Base stock	Partial

businesses are flooded with requests for a particular item, it is natural at some point to offer customers an alternative item due to excessive waiting times. Partial backordering can be a useful policy for preventing customers from reneging. Customers may renege for various reasons, for example, if they decide to rent or purchase the item from another company or decide to substitute it by another item. Reneging leads to unreliable estimates of waiting times since customers often tend not to notify the rental business when they no longer require the item. Since it takes time to notice that items put aside for a customer are not retrieved, items are also unnecessarily out of circulation, causing significant unavailability costs.

Our contributions are as follows. We determine optimal and near-optimal base stock levels for rental systems with partial backordering. In this setting, we are the first to fully characterize the optimal base stock policy for a serial system, i.e., a system with one local warehouse and one support warehouse, which models a variety of interesting practical applications. We propose a new greedy algorithm based on an approximated cost function for given base stock levels. This algorithm yields near-optimal base stock levels and outperforms existing methods for no backordering, showing the importance of accounting for partial backordering. In addition, we contribute to the current literature by obtaining several managerial insights into the reduction of lost demand due to partial backordering and the benefit of having a support warehouse with lower holding costs.

The paper is organized as follows. First we review relevant literature in Section 2. In Section 3 we formulate the model and evaluate steady-state costs of given base stock policies. We characterize optimal policies for the serial system in Section 4. Based on this special case, we propose upper bounds on base stock levels in the general system in Section 5. In order to obtain solutions in large instances, we propose an efficient approximation of the steady-costs in Section 6 and a greedy algorithm in Section 7. Numerical experiments are carried out in Section 8 to demonstrate the performance of the greedy algorithm and to obtain managerial insights. Finally, Section 9 concludes.

2. Literature review

We now review literature related to our rental system with expedited shipments from a support warehouse. We first discuss literature about support warehouses of which an overview is given in Table 1. Afterwards, we comment on the relation with lateral transshipment and two-echelon models and discuss partial backordering in other inventory systems.

The existing literature on support warehouses in Table 1 concerns inventory control for consumable spare parts. An important stream of literature, starting with the seminal work of Sherbrooke (1968), deals with repairable spare parts. To the best of our knowledge support warehouses have not been considered for repairable spare parts. However, when base stock policies are assumed, the analysis for consumable and repairable spare parts is generally the same.

In spare parts systems with support warehouses, inventory at each warehouse is typically replenished using one-for-one base

stock policies. This is the same as in a rental system: the base stock level corresponds to the number of rental items, and replenishment lead times to rental durations. A key difference is the backordering process. Due to service arrangements with clients, spare parts providers typically have full backordering or no backordering, whereas rental systems typically have partial backordering. Another difference is the number of rented items/orders when backordering is possible. Items can only be rented if there is on-hand stock, so in rental systems the number of rented items is inherently limited by the base stock level. For spare parts, each demand generates a new order, so the number of orders exceeds the base stock level when there are backorders. For the above reasons, existing spare parts models cannot be straightforwardly applied to our rental system.

We now discuss in more detail differences between existing models and our work. Kranenburg and Van Houtum (2009) consider multiple support warehouses with no backordering. Near-optimal base stock levels are obtained using a greedy heuristic and an accurate approximation for the cost function of given base stock policies. We instead consider a single support warehouse, however, we extend their model with partial backordering and propose a new approximation of the cost function for this setting. Furthermore, we complement their work by providing upper bounds on base stock levels and proving the structure of the optimal base stock policy for serial systems. In a system with a single support warehouse, Howard, Marklund, Tan, and Reijnen (2015) introduce a threshold backorder policy. Demand at a local warehouse is backordered if an order will be delivered within a specified threshold time, otherwise it is met with an expedited shipment from the support warehouse. Even though threshold backordering is closely related to our setting, important differences remain. A threshold rule may be appropriate when delivery times of spare part suppliers are reliable, but less so with uncertain return times of rented items. In our rental system, expedited shipments are preferred to letting customers wait, hence demand is backordered only when expedited shipments are not possible.

The remaining contributions in Table 1 focus on different policies. Van Wijk, Adan, and Van Houtum (2013) optimize the policy of accepting expedited shipment requests at the support warehouse in case of varying penalty costs at local warehouses. Assuming all backorders are placed at the support warehouse, Axsäter, Howard, and Marklund (2013) optimize batch-ordering policies in a setting with normal demand.

Models with support warehouses are related to lateral transshipments and two-echelon distribution systems. For example, an expedited shipment from a support warehouse can be regarded as a unidirectional lateral transshipment (Axsäter, 2003). Lateral transshipment research typically focuses on situations with either full backordering or lost sales in single-echelon settings (Paterson, Kiesmüller, Teunter, & Glazebrook, 2011). Because expedited shipments originate from one warehouse, there is overlap with two-echelon distribution systems. Basten and Van Houtum (2013) consider a two-echelon distribution system with base stock control, Poisson demand, and full backordering. The authors provide new properties of the steady-state costs and formulate a smart enumeration procedure and a heuristic to obtain near-optimal base stock levels. Rong, Atan, and Snyder (2014) propose a heuristic for obtaining base stock policies in multi-echelon distribution systems by decomposing it into serial systems. We apply a similar idea by using optimal base stock policies of serial systems as bounds for base stock levels in the general system.

Various inventory control models with partial backordering are reviewed in Bijvank and Vis (2011). Frequently, a specified fraction of customers is backordered while the other customers are lost (Montgomery, Bazaraa, & Keswani, 1973). Several authors set a partial backorder level as in our work. For a single warehouse

Fig. 1. A rental system with a support warehouse and n local warehouses.

with Poisson demand and (R, Q) policies, Rabinowitz, Mehrez, Chu, and Patuwo (1995) show that, for certain ranges of the backorder cost per time unit, partial backordering yields significant cost savings over no and full backordering. Essentially, this same result is shown by Thangam and Uthayakumar (2008) for a two-echelon system with Poisson demand and (R, Q) policies. Chu, Patuwo, Mehrez, and Rabinowitz (2001) show that cost savings are further improved by setting different partial backorder levels during different time periods. Given its relevance in these related settings, we believe it is important to investigate the added value of partial backordering in support warehouse models.

3. Model and assumptions

In this section, we present the model and assumptions for rental systems with several local warehouses (LW) and expedited shipments from a support warehouse (SW). Since there are no other shipments in our model, from now on we refer to expedited shipment as shipment. We subsequently model the system as a Markov chain and derive its steady-state costs. The notation used is similar to that of Basten and Van Houtum (2013) and Kranenburg and Van Houtum (2009).

We consider a single item rental system with one support warehouse and n local warehouses. The SW is indexed by $i = 0$ and the n LWs are indexed by $i = 1, \dots, n$. We refer to the total number of rental items at warehouse i as base stock level S_i , $i = 0, \dots, n$. The base stock policy for the system is given by the vector of base stock levels $S = (S_0, S_1, \dots, S_n)$. The system is shown schematically in Fig. 1. The arrows denote the possible shipments from the SW in case of stock-outs at the LWs.

Demand at LW i , $i = 1, \dots, n$ arrives according to a stationary independent Poisson process with rate λ_i . The SW faces no external demand. In case of a stock-out at an LW, the SW immediately carries out a shipment to meet the demand, provided stock is available at the SW. The lead time for shipments is zero. Shipments typically take place on the same day as the demand, which is small compared to the rental duration of several days or weeks. Even if the lead time is small, we can assume it to be zero and then include costs for extra unavailability in the shipment costs.

Each item used for fulfilling demand is rented for a stochastic rental time and returned afterwards. We assume that items originating from the SW are returned to the SW after the rental period, because the SW has the ownership of the shipped item. For all warehouses, the rental time is identically and independently exponentially distributed with mean μ^{-1} . We select an identical distribution because the rental time is typically product dependent and not warehouse dependent. Similar to Alfredsson and Verrijdt (1999), an experiment in Section 8.2 indicates that optimal base stock levels are insensitive to the choice of rental time distribution, provided the mean is the same. Assuming exponential rental times is therefore not restrictive.

There is partial backordering. When the SW is out of stock, up to R demands per LW can be backordered; additional demand is lost. Lost demands typically mean that customers either rent a substitute product or go to a competitor. We assume $0 \leq R < \infty$. The $R = 0$ case is similar to that considered by Kranenburg and Van Houtum (2009). The $R = \infty$ case is computationally intractable,

however, it can be approximated by setting R sufficiently high so that virtually no backorders are blocked. Backorders at an LW can either be met by a returned item at that LW or by a shipment if an item returns at the SW. In case there are multiple LWs with backorders, the item returned at the SW is assumed to be shipped randomly to one of these LWs, with a probability equal to the LW's relative share of the total number of backorders in the system. In other words, each customer with a backorder has an equal probability of receiving a shipment.

The SW has holding costs $h_0 > 0$ per unit time and the LWs have identical holding costs $h_i = h > 0$. As the SW is dedicated to storage, it is assumed that $h_0 \leq h$. The shipment cost is $c > 0$ per shipped demand. The shipment cost is identical for all LWs and includes the cost of returning the item to the SW after renting. There is a one-time backorder cost $b > 0$ for each backordered demand and lost demand cost $\ell > 0$ for each lost demand. In most practical settings, we see that $\ell \geq b$ because customers typically face more dissatisfaction from a lost demand than from a backorder. For our rental system we make the stronger assumption that $\ell \geq b + c$, so that meeting backorders by shipments is preferred over losing demand. In addition, we assume $b \geq c$ so that shipments are worthwhile. These assumptions on the cost parameters are reasonable because these cost incentives are the main motivation to consider shipments from an SW in the first place.

We have several remarks regarding identical model parameters and the choice of a pure cost model. Identical costs at LWs remove incentives for rationing stock at the SW; we refer to Van Wijk et al. (2013) for optimal rationing policies in settings with non-identical costs. Technical implications of identical rental times will be explained at a more appropriate point during the analysis in Section 4.

Our model is a pure cost model. In case backorder and lost demand costs prove challenging to estimate in practice, one can instead impose service level constraints. The easiest way to incorporate service levels is by carrying out a two-dimensional bisection on ℓ and b until suitable service levels are reached. Alternatively, one can set $\ell = b = 0$ in all cost functions and adapt optimization procedures to satisfy the service level constraint. The latter approach is tested in Section 8.3.

3.1. Markov chain

We model the rental system as a continuous time Markov chain and derive the stationary distribution and steady-state costs for a given base stock policy $S = (S_0, \dots, S_n)$.

Let x_i be the current available stock at warehouse i , $i = 0, \dots, n$, which is negative in case of backorders. Then a state x is represented by

$$x = (x_0, x_1, \dots, x_n),$$

with

$$x_0 \in \{0, \dots, S_0\} \text{ and } x_i \in \{-R, \dots, S_i\} \text{ for } i = 1, \dots, n.$$

Since the SW always ships when an LW has backorders, states x with simultaneously $x_0 > 0$ and $-R \leq x_i < 0$ for some i cannot occur.

Define e_i as the vector with a 1 at index i and 0 at all other indices, so that if in state x an item returns at LW i , the new state becomes $x + e_i$. Furthermore, let $x_i^+ = \max\{x_i, 0\}$ and $x_i^- = \max\{-x_i, 0\}$, and let $\mathbb{1}\{E\}$ be the indicator function with $\mathbb{1}\{E\} = 1$ if E is true and 0 otherwise.

We state the transition rates first and then discuss them. The transition rates are

$$q(x, x - e_i) = \lambda_i \mathbb{1}\{x_i > 0\} + \lambda_i \mathbb{1}\{x_0 = 0, -R < x_i \leq 0\}, \quad (1)$$

$$q(x, x - e_0) = \sum_{i=1}^n \lambda_i \mathbb{1}\{x_0 > 0, x_i = 0\}, \quad (2)$$

$$q(x, x + e_i) = \mu(S_i - x_i)^+ + \mu S_0 r_i(x) \mathbb{1}\{x_i < 0\}, \quad (3)$$

$$q(x, x + e_0) = \mu(S_0 - x_0) \mathbb{1}\{x_i \geq 0 \text{ for } i = 1, \dots, n\}, \quad (4)$$

with

$$r_i(x) = \frac{x_i^-}{\sum_{j=1}^n x_j^-}.$$

Eq. (1) gives the rate by which stock at LW i decreases, $i = 1, \dots, n$. Demand is either met from on-hand stock or backordered when both the SW and LW i are out of stock while LW i is not yet at its maximum partial backorder level. Eq. (2) shows the rate by which items are shipped from the SW, i.e., the combined demand rates of LWs with $x_i = 0$. Eq. (3) concerns increases in stock at LW i due to returns. Items return at LW i with rate $\mu(S_i - x_i)^+$ and are shipped from the SW with probability $r_i(x)$ in case LW i has a backorder and an item returns at the SW. $r_i(x)$ is the relative share of LW i in the total number of backorders in the system. Finally, in Eq. (4) items return at the SW with rate $\mu(S_0 - x_0)$ only if no LWs have backorders.

3.2. Steady-state costs

Using standard theory for continuous time Markov chains, we obtain the steady-state probabilities $\pi(x)$ for each state x (see, e.g., Bolch, Greiner, De Meer, and Trivedi, 2006, p. 96). The steady-state costs follow from the PASTA property. In each state with $x_0 > 0$ and $x_i = 0$ shipment costs c are incurred at rate λ_i . Moreover, in each state with $x_0 = 0$ and some $x_i < 0$, shipments are carried out with rate $S_0\mu$ if an item returns at the SW. So the shipment costs for base stock policy S are

$$\begin{aligned} \bar{c}(S) = c \sum_x \pi(x) & \left(\sum_{i=1}^n \lambda_i \mathbb{1}\{x_0 > 0, x_i = 0\} \right. \\ & \left. + S_0\mu \mathbb{1}\{x_0 = 0, \exists i : x_i < 0\} \right). \end{aligned} \quad (5)$$

With holding costs h_i per time unit, the holding costs for both support and local warehouse i , $i = 0, \dots, n$ are

$$\bar{h}_i(S) = h_i \sum_x \pi(x) x_i^+.$$

With cost b per backorder and ℓ per lost demand, the total backorder costs for LW i , $i = 1, \dots, n$ are

$$\bar{b}_i(S) = b \lambda_i \sum_x \pi(x) \mathbb{1}\{x_0 = 0, -R < x_i \leq 0\},$$

and lost demand costs

$$\bar{\ell}_i(S) = \ell \lambda_i \sum_x \pi(x) \mathbb{1}\{x_0 = 0, x_i = -R\}.$$

The steady-state costs per time unit are given by

$$C(S) = \bar{c}(S) + \bar{h}_0(S) + \sum_{i=1}^n (\bar{h}_i(S) + \bar{b}_i(S) + \bar{\ell}_i(S)).$$

Note that $C(S)$ and $\pi(x)$ can be obtained similarly even if μ , c , b , and ℓ are not identical for all LWs.

Our objective is to find the cost-minimizing vector of base stock levels S^* , that is,

$$S^* = \arg \min_S C(S).$$

This optimization problem is more difficult to solve than the $R = 0$ case considered in Kranenburg and Van Houtum (2009), since the state space for a given S has $\prod_{i=1}^n (S_i + R + 1) - \prod_{i=1}^n (S_i + 1)$ extra states. Provided we have upper bounds on the optimal base stock levels, we can, in principle, enumerate all possible solutions. We derive these upper bounds in Section 5, based on an analysis of special cases in Section 4. As calculating $C(S)$ is not tractable for large instances, in Section 6 we propose an easy-to-compute approximation for $C(S)$, and as enumerating is exhaustive, we propose a more efficient procedure in Section 7.

4. Serial system

In this section we determine the optimal base stock policy for a serial system, i.e., a system with one SW and one LW. This notion of serial system differs from its traditional notion in multi-echelon models, however, we call it that since two warehouses are ordered in series for fulfilling demand. This special case models various relevant practical applications, for example warehouses or university libraries with a front area for demand fulfillment and a back area for reserve storage.

We derive an exact expression for the steady-state costs of a given base stock policy. We show that the optimal policy structure can be characterized by a trade-off between shipment and holding costs. That is, there is a threshold base stock level at the LW beyond which we are willing to store items at the SW to incur the low holding cost h_0 (instead of h) at the expense of costs for occasional shipments. Using this threshold structure, the optimal base stock levels follow from a trade-off between the holding and shipment costs and the backorder and lost demand costs.

Throughout the paper we apply results from the $M/M/s/s$ loss queue and $M/M/s/s + k$ finite queue. For the $M/M/s/s$ loss queue with s servers, we use the following well-known quantities, see also Öner, Kiesmüller, and Van Houtum (2009). Let $a = \frac{\lambda}{\mu}$. The blocking probability is given by

$$B(s, a) = \frac{a^s / s!}{\sum_{k=0}^s a^k / k!},$$

the carried load by

$$A(s, a) = a(1 - B(s, a)), \quad (6)$$

and the load carried by the last server by

$$F(s, a) = a(B(s - 1, a) - B(s, a)).$$

For the $M/M/s/s + k$ queue, we require the following steady-state probabilities, which can be obtained numerically using the recursion for the steady-state distribution in Gross, Shortle, Thompson, and Harris (2008, Chapter 2). Letting x denote the number of available servers, taken negative in case of delayed customers, we define

$$B^k(s, a) = P(x = -k), \quad (7)$$

$$W^k(s, a) = P(-k < x \leq 0),$$

$$P_k \equiv P^k(s, a) = P(-k \leq x < 0),$$

$$A^k(s, a) = a(1 - B^k(s, a)), \quad (8)$$

as the blocking probability, delay probability, probability of seeing delayed customers, and carried load, respectively. It is useful to observe that for $k = 0$, $B^k(s, a) = B(s, a)$, $W^k(s, a) = 0$, $P_k = 0$, and $A^k(s, a) = A(s, a)$.

4.1. Steady-state costs

We now derive an expression for the steady-state costs in the serial system for a given base stock policy $S = (S_0, S_1)$ and partial

Fig. 2. Transition graph for the serial system with $S_0 = 2, S_1 = 2, R = 1$. The state (x_0, x_1) gives the stock levels at the SW and LW.

backorder level R . We characterize the steady-state distribution in Lemma 1 and use this to formulate the costs in Theorem 1. Fig. 2 shows an example transition graph of the Markov chain that may aid in understanding the proof of Lemma 1.

Let $\bar{S} = S_0 + S_1$ be the total stock. Based on queueing models, we specify several aggregate steady-state probabilities.

Lemma 1. Aggregate steady-state probabilities for the serial system.

- For $k = -R, -R+1, \dots, \bar{S}$, the aggregate steady-state probability $P(x_0 + x_1 = k)$ equals the steady-state probability of k available servers in the $M/M/\bar{S}/\bar{S}+R$ queue.
- For $k = 0, 1, \dots, S_1$, the aggregate steady-state probability $P(x_1 = k)$ equals the steady-state probability of x_1 available servers in an $M/M/S_1/S_1$ queue, scaled by $1 - P_R$.

Proof.

- The aggregate stock $x_0 + x_1$ decreases by 1 with rate λ_1 if $x_0 + x_1 > -R$ and increases by 1 with rate $\mu(\bar{S} - x_0 - x_1)$ if $x_0 + x_1 \geq 0$ and with rate $\mu\bar{S}$ if $x_0 + x_1 < 0$ (refer to Fig. 2). These are exactly the transition rates of the $M/M/\bar{S}/\bar{S}+R$ queue.
- The aggregate transition rates between states with $x_1 \geq 0$ equal those of the $M/M/S_1/S_1$ loss queue (refer to Fig. 2). By balance of flow, the relative frequency with which we observe stock levels $x_1 \geq 0$ must be the same as observing x_1 available servers in the $M/M/S_1/S_1$ queue. The total probability of observing $x_1 \geq 0$ equals $1 - P_R$, hence we need to scale by $1 - P_R$. $\square$

Even though the steady-state probabilities are known only on aggregate, this is in fact sufficient to derive the steady-state number of shipments, backorders, and lost demands, and the carried load at each warehouse. Therefore we can now specify the costs in Theorem 1. In Section 4.2 we analyze these costs to prove the structure of the optimal base stock policy.

Theorem 1. Letting $a_1 = \frac{\lambda_1}{\mu}$, the steady-state costs of the serial system are given by

$$C(S_0, S_1) = \bar{c}(S_0, S_1) + \bar{h}_0(S_0, S_1) + \bar{h}_1(S_0, S_1) + \bar{b}_1(S_0, S_1) + \bar{\ell}_1(S_0, S_1), \quad (9)$$

with

$$\begin{aligned} \bar{c}(S_0, S_1) &= c\lambda_1((1 - P_R)B(S_1, a_1) + P_R - B^R(\bar{S}, a_1) - W^R(\bar{S}, a_1)) \\ &\quad + c\mu S_0 P_R, \end{aligned}$$

$$\bar{h}_1(S_0, S_1) = h(1 - P_R)(S_1 - A(S_1, a_1)),$$

$$\bar{h}_0(S_0, S_1) = h_0(S_0 - A^R(\bar{S}, a_1) + (1 - P_R)A(S_1, a_1) + P_R S_1),$$

$$\bar{b}_1(S_0, S_1) = b\lambda_1 W^R(\bar{S}, a_1),$$

$$\bar{\ell}_1(S_0, S_1) = \ell\lambda_1 B^R(\bar{S}, a_1).$$

Proof. See appendix. $\square$

The expressions in Theorem 1 require identical μ 's at the SW and LW. In case of nonidentical μ 's, the serial system does not reduce to well-known one-dimensional queues and we can only obtain costs numerically.

Since the goal is to minimize costs, we have a useful convexity result for the special case $R = 0$.

Lemma 2. For $R = 0$, the steady-state costs $C(S_0, S_1)$ are convex in both S_0 and S_1 .

Proof. For $R = 0$, Eq. (9) can be rewritten as

$$\begin{aligned} C(S_0, S_1) &= h_0 S_0 + h S_1 - h a_1 + (\mu(\ell - c) + h_0) a_1 B(\bar{S}, a_1) \\ &\quad + (c\mu + h - h_0) a_1 B(S_1, a_1), \end{aligned}$$

which is convex in both S_0 and S_1 because $B(s, a)$ is convex in s (Messlerli, 1972). $\square$

By plotting, it can be observed that $C(S_0, S_1)$ is generally not convex for $R > 0$. However, it appears to be quasiconvex in S_0 and S_1 provided $\ell \geq b + c$. We have been unable to formally prove this quasiconvexity, though it seems intuitive under the assumed cost structure.

4.2. Threshold structure of the optimal base stock policy

We now characterize the optimal trade-off between holding and shipment costs in the serial system. We show that an extra item is stored cheapest at the LW below a certain threshold base stock level and at the SW otherwise. Whereas Wee and Dada (2005) show such a threshold in single period problem, we now show this for the situation with continuous review and partial backordering.

In order to show this threshold, we consider the marginal difference in costs between adding an extra item to the LW or the SW. For arbitrary base stock levels S_0 and S_1 , define the marginal difference as

$$\Delta C^R(S_0, S_1) = C(S_0, S_1 + 1) - C(S_0 + 1, S_1). \quad (10)$$

If $\Delta C^R(S_0, S_1) < 0$ then the extra item is stored cheapest at the LW, otherwise at the SW. We derive $\Delta C^R(S_0, S_1)$ below.

Lemma 3.

$$\Delta C^R(S_0, S_1) = h - h_0 - (c\mu + h - h_0)(F(S_1 + 1, a_1)(1 - P_R) + P_R). \quad (11)$$

Proof. Substitute expressions for $C(S_0, S_1)$ from Theorem 1 into Eq. (10) and rearrange terms. $\square$

Note that costs for lost and backordered demand do not appear in the marginal difference since total stock equals $S_0 + S_1 + 1$ in both situations.

Now we prove several properties of $\Delta C^R(S_0, S_1)$.

Lemma 4. Properties of $\Delta C^R(S_0, S_1)$.

- $\Delta C^R(S_0, S_1)$ increases in S_1 for every fixed S_0 and increases in S_0 for every fixed S_1 .
- For fixed S_0 , $\lim_{S_1 \rightarrow \infty} \Delta C^R(S_0, S_1) = h - h_0$.

Proof.

- (i) First fix S_0 . Since $(c\mu + h - h_0) > 0$, we need to show $F(S_1 + 1, a_1)(1 - P_R) + P_R$ decreases in S_1 . The probability P_R of observing delayed customers clearly decreases in S_1 . By Öner et al. (2009), $F(S_1 + 1, a_1)$ decreases in S_1 . Since $F(S_1 + 1, a_1) \leq 1$, P_R decreases faster than $F(S_1 + 1, a_1)(1 - P_R)$ can increase. Hence $F(S_1 + 1, a_1)(1 - P_R) + P_R$ decreases in S_1 and the result follows. Now fix S_1 . $\Delta C^R(S_0, S_1)$ depends on S_0 only through P_R . Since P_R decreases in S_0 , $\Delta C^R(S_0, S_1)$ increases in S_0 .
- (ii) This follows since $\lim_{S_1 \rightarrow \infty} F(S_1 + 1, a_1)(1 - P_R) + P_R = 0$. $\square$

The properties from Lemma 4 imply a threshold policy on base stock level S_1 at the LW. For a given S_0 , we can increase S_1 until a certain threshold value $S'_{1,R}(S_0)$ where $\Delta C^R(S_0, S_1)$ becomes strictly positive (unless $h = h_0$). Theorem 2 summarizes these threshold results by checking the values of S_0 and S_1 for which $\Delta C^R(S_0, S_1) \leq 0$ and $\Delta C^R(S_0, S_1) > 0$.

Theorem 2. Threshold results for the serial system.

- (i) The optimal policy for the serial system has a threshold structure.
- (ii) For fixed S_0 , a threshold level $S'_{1,R}(S_0)$ on S_1 is given by
- $$S'_{1,R}(S_0) = \min\{S_1 : \Delta C^R(S_0, S_1) > 0\}. \quad (12)$$
- (iii) Threshold $S'_{1,R}(S_0)$ has bounds $S'_{1,0}(0) \leq S'_{1,R}(S_0) \leq S'_{1,R}(0)$.
- (iv) If $h_0 = h$, then $S'_{1,R}(S_0) = \infty$.

Proof.

- (i–ii) Immediate from Lemma 4.
- (iii) Rewriting $\Delta C^R(S_0, S_1)$ in terms of $\Delta C^0(0, S_1)$ yields
- $$\Delta C^R(S_0, S_1) = \Delta C^0(0, S_1) - (c\mu + h - h_0)P_R(1 - F(S_1 + 1, a_1)).$$
- This shows $\Delta C^R(S_0, S_1) \leq \Delta C^0(0, S_1)$, hence $S'_{1,0}(0) \leq S'_{1,R}(S_0)$. Moreover, since for every fixed S_1 , $\Delta C^R(S_0, S_1)$ increases in S_0 , clearly $\Delta C^R(0, S_1) \leq \Delta C^R(S_0, S_1)$. Hence, $S'_{1,R}(S_0) \leq S'_{1,R}(0)$.
- (iv) For $h_0 = h$, $\Delta C^R(S_0, S_1) \leq 0$ for all S_0 and S_1 , so the threshold is ∞ . $\square$

Note that for $R = 0$, the lower and upper bound on $S'_{1,R}(S_0)$ both equal $S'_{1,0}(0)$, so the threshold is constant. For $R > 0$, the threshold is constant only if $S'_{1,0}(0) = S'_{1,R}(0)$. In the special case $h_0 = h$, evidently no stock should be stored at the SW.

4.3. Procedure for determining base stock levels

We now give a procedure for determining base stock levels in the serial system, using the fact that an optimal base stock policy satisfies the threshold structure from Theorem 2. In essence, the threshold structure tells us how to divide a given number of items between the LW and SW, based only on holding and shipment costs. In order to determine the optimal base stock levels we also need to take into account costs for backordered and lost demand.

The procedure is as follows.

Serial system procedure:

1. Calculate $S'_{1,0}(0)$ and $S'_{1,R}(0)$ using Eq. (12).
2. Start at $S_0 = 0, S_1 = 0$. While $S_1 < S'_{1,R}(S_0)$, increase S_1 by 1. If $C(0, S_1) > C(0, S_1 - 1)$, then stop and return $S_0^* = 0, S_1^* = S_1 - 1$ as optimal solution. Continue to step 3 if this condition is not met for any $S_1 \leq S'_{1,R}(0)$.
3. Repeatedly increase S_0 by 1 until $C(S_0, S'_{1,R}(S_0)) > C(S_0 - 1, S'_{1,R}(S_0 - 1))$. Return $S_0^* = S_0 - 1, S_1^* = S'_{1,R}(S_0 - 1)$ as optimal solution.

Fig. 3. Example of the serial system procedure.

Fig. 3 shows an example of the procedure. While costs decrease, we increase S_1 until the threshold is reached. If no optimum is found, we increase S_0 , determine its corresponding threshold value, and check the costs of that policy until costs start to increase. In this way, all solutions satisfying the threshold structure are checked.

The procedure is exact for $R = 0$ due to the convexity of $C(S_0, S_1)$ established in Lemma 2. The procedure is exact for $R > 0$ if $C(S_0, S_1)$ is quasiconvex. If this is not the case, one can increase S_0 until $|C(S_0, S_1) - C(S_0 - 1, S_1) - h_0| < \varepsilon$ with $\varepsilon > 0$ a small number, since in the limit $C(S_0, S_1)$ increases strictly in S_0 with slope h_0 .

4.4. Optimal base stock policy in a decoupled system

The serial system results can be directly applied to obtain the optimal base stock policy in the important special case with $S_0 = 0$ and an arbitrary number of LWs. This special case has n decoupled LWs that independently meet their demand without shipments from the SW. Let S_i^d denote the optimal decoupled base stock level for LW i , $i = 1, \dots, n$. Clearly, the steady-state costs for LW i are given by the costs of the serial system evaluated in $S_0 = 0$. We can therefore obtain S_i^d by minimizing $C(0, S_i)$ in S_i for each $i = 1, \dots, n$.

5. Upper bounds for optimal base stock levels

Now we state upper bounds on S^* for the $R = 0$ case in the general system, i.e., the system with any number of LWs. These upper bounds limit the number of base stock policies that we need to enumerate in an optimization procedure.

First we state bounds for S_i , $i = 1, \dots, n$. Let S_i^S be the optimal base stock level at LW i in the serial system. We now show that $S_i^* \leq S_i^S$.

Lemma 5. In the general system with $R = 0$, there exists an optimal solution S^* with $S_i^* \leq S_i^S$ for all $i = 1, \dots, n$.

Proof. Suppose there is an optimal base stock policy S^* with $S_i^* > S_i^S$ for some specific LW i , $1 \leq i \leq n$. Call this System I. Partition S_i^* into $S_{i,0}$ and $S_{i,1}$ so that $S_i^* = S_{i,0} + S_{i,1}$. In System II, take the same base stock policy as System I, but store $S_{i,0}$ items at the SW as dedicated stock for shipments to LW i only, shipped before using the other S_0^* items at the SW. Evidently, Systems I and II differ only in the stock process for items $S_{i,0}$ and $S_{i,1}$ and are identical when $S_{i,1} = S_i^*$. For $R = 0$, this stock process is a serial system with a lost demand cost that is either c or ℓ depending on whether the SW has stock. Since S_i^S is optimal for a serial system with higher or equal lost demand cost, we must have that an optimal $S_{i,1}^* \leq S_i^S$. System II therefore has lower costs than System I.

In System III, take the same base stock policy as in System II, but share the dedicated $S_i^* - S_{i,1}^*$ items at the SW for LW i with all

other LWs. The stock processes at the LWs are identical in Systems II and III ($M/M/s/s$ queues with the same s), hence the SW faces the same shipment requests. For any sample path of shipment requests, the number of lost demands and the number of on-hand items in System III are lower than or equal to the same quantities in System II. System III therefore has lower costs than System II, and thus lower costs than System I. This shows there exists an optimal solution with $S_i^* \leq S_i^S$ for all $i = 1, \dots, n$. $\square$

The method used in the above proof does not apply for $R > 0$ since the stock processes at the LWs are not serial systems for $R > 0$ unless $n = 1$. However, our numerical experiments in Section 8.1 indicate that use of S_i^S leads to near-optimal results for $R > 0$.

As upper bound for S_0^* , we let the SW deal with all the demand in the system. We set $S_i = 0$ at all LWs and then determine the cost-minimizing number of items at the SW.

Lemma 6. For the $R = 0$ case, an upper bound for S_0^* is given by

$$S_0^u = \max\{S_0 : \mu(\ell - c)F(S_0, a_0) > h_0(1 - F(S_0, a_0))\},$$

with $a_0 = \frac{\lambda_0}{\mu}$ and $\lambda_0 = \sum_{i=1}^n \lambda_i$.

Proof. When $S_i = 0$ for all $i = 1, \dots, n$, the on-hand stock at the SW is an $M/M/S_0/S_0$ queue with arrival rate λ_0 . Therefore, its costs are given by $C(S_0, 0)$ of the serial system with demand rate λ_0 . These costs are convex in S_0 by Lemma 2, so, starting at $S_0 = 0$, S_0^u follows by checking when $C(S_0 - 1, 0) < C(S_0, 0)$ for the first time.

For S_0^u to be an upper bound, it is necessary that $S_0^*(S_1, \dots, S_n)$ is decreasing (i.e., non-increasing) in each S_i , $i = 1, \dots, n$, where we define $S_0^*(S_1, \dots, S_n)$ as the optimal S_0 for given base stock levels S_i , $i = 1, \dots, n$. Now increase S_i of some LW i , $1 \leq i \leq n$. Take any sample path of demands. For this sample path, the number of shipment requests at the SW is lower than or the same as before. Hence, each item at the SW prevents less lost demand and is on-hand more often. The marginal costs for each item at the SW increase, so $S_0^*(S_1, \dots, S_n)$ decreases in S_i . $\square$

An upper bound for $R > 0$ can be obtained by solving the Markov chain for different values of S_0 and picking the cost-minimizing S_0 , however, this is intractable for large systems. Since one of our goals is to determine good solutions in large systems, in Section 6 we propose an approximate cost function that has a bound that can efficiently be calculated for $R > 0$.

6. Approximated cost function

Calculation of the exact steady-state costs from Section 3.2 is intractable for systems with many LWs due to the state space size. We therefore formulate an easy-to-compute approximation for the steady-state costs by decoupling the warehouses. Such decoupling approaches have been shown to perform well in other multi-location settings with base stock control (Alfredsson & Verrijdt, 1999; Axsäter, 1990; Kukreja, Schmidt, & Miller, 2001).

We extend the approximation of Kranenburg and Van Houtum (2009) to address situations with partial backordering. The system is decoupled into $n + 1$ queues. The n LWs are modeled as $M/M/s/s$ queues with $s = S_i$. The blocked customers in these queues are redirected towards the SW from which they receive an item through shipment. The SW is modeled as an $M/M/s/s + k$ queue with $s = S_0$ and $k = nR$. Arriving customers wait in queue when the SW has no available items and are rejected when there are k customers in the queue. The approximation has a partial backorder level nR for the whole system instead of a partial backorder level R per LW. Although apparently different, experiments in Section 8.1 indicate that the approximation performs well for obtaining near-optimal base stock levels.

The arrival rate λ_0 for the $M/M/s/s + k$ queue of the SW is given by the total demand blocked at the LWs, i.e.,

$$\lambda_0 = \sum_{i=1}^n \lambda_i B(S_i, a_i).$$

We now give expressions for the steady-state costs of the approximation in terms of the blocking and delaying probabilities from Eqs. (7) and (8), similar as in Theorem 1. We approximate $C(S)$ by

$$\tilde{C}(S) = \tilde{c}(S) + \tilde{b}(S) + \tilde{\ell}(S) + \sum_{i=0}^n \tilde{h}_i(S), \quad (13)$$

where

$$\tilde{c}(S) = c\lambda_0(1 - B^{nR}(S_0, a_0)),$$

$$\tilde{b}(S) = b\lambda_0 W^{nR}(S_0, a_0),$$

$$\tilde{\ell}(S) = \ell\lambda_0 B^{nR}(S_0, a_0),$$

$$\tilde{h}_0(S) = h_0(S_0 - A^{nR}(S_0, a_0)),$$

$$\tilde{h}_i(S) = h(S_i - A(S_i, a_i)).$$

6.1. Upper bound for base stock level at the SW

For the upper bound S_0^u at the SW under the approximated cost function, we apply the same idea as in Section 5. Again we set $S_i = 0$ for $i = 1, \dots, n$ and let the SW deal with all demand. In fact, this is a special case of the serial system with partial backorder level nR and demand rate λ_0 . Therefore we can minimize $C(S_0, 0)$ in S_0 to obtain the upper bound S_0^u .

6.2. Properties of the approximated cost function

We now prove properties of the approximated cost function that simplify the search for the base stock policy with the lowest approximate costs. First we show how we can determine the cost-minimizing base stock levels $S_i^*(S_0)$ for $i = 1, \dots, n$. Starting from an arbitrary base stock policy, in Lemma 7 we determine the LW for which reducing its base stock level by one leads to the largest cost reduction.

Lemma 7. Let $S = (S_0, S_1, \dots, S_n)$ be a base stock policy such that $I = \{i : S_i > 0, i \geq 1\} \neq \emptyset$. Suppose we must reduce S to $S - e_i$ for some $i \in I$. The approximate cost function is minimized by selecting LW i^* with

$$i^* = \underset{i \in I}{\operatorname{argmin}} F(S_i, a_i).$$

Proof. See appendix. $\square$

Lemma 7 implies that if for some fixed S_0 we start with high enough values of S_i , $i = 1, \dots, n$ then greedily reducing base stock levels of the LWs yields $S_i^*(S_0)$. The optimal rule is to subtract from the LW with the smallest $F(S_i, a_i)$. As $F(S_i, a_i)$ depends only on the local base stock level S_i , the optimal sequence of removing items from the LWs remains unchanged when S_0 changes.

Another useful property is $S_i^*(S_0 + 1) \leq S_i^*(S_0)$, i.e., the optimal base stock levels at an LW are decreasing in S_0 . If this is the case, then solution $S_i^*(S_0)$ can serve as starting solution for determining $S_i^*(S_0 + 1)$. We show indeed that this property holds for $R = 0$ in Lemma 8. The proof essentially shows that the cost reduction for removing an item, i.e., $\tilde{C}(S) - \tilde{C}(S - e_i)$, increases in S_0 .

Lemma 8. For $R = 0$, the optimal base stock levels under the approximated cost function satisfy $S_i^*(S_0 + 1) \leq S_i^*(S_0)$.

Proof. See appendix. $\square$

For $R > 0$ a similar proof does not apply since $\tilde{C}(S) - \tilde{C}(S - e_i)$ is in general not monotonically increasing. However, we conjecture that the property $S_i^*(S_0 + 1) \leq S_i^*(S_0)$ also holds for $R > 0$ since we were unable to find counter examples in thousands of randomly generated instances.

7. Greedy algorithm

In this section, we develop an efficient algorithm for determining the base stock policy with minimal approximate costs. In numerical experiments we observe that the costs $\tilde{C}(S_0, S_1^*(S_0), \dots, S_n^*(S_0))$ are not convex in S_0 . The idea for the algorithm is therefore to calculate the cost-minimizing base stock levels $S_i^*(S_0)$ for all values of S_0 between certain bounds. These solutions are compared with each other to determine the global cost-minimizing base stock levels. Although upper bounds S_i^* in general do not apply for the approximated cost function, we use S_i^* as initial solution because the optimal policy that we aim to approximate is located in that region. Using Lemma 8, we continue each subsequent iteration with the previously found solution to reduce the computational effort. The procedure is as follows.

Greedy algorithm:

1. Determine the decoupled solution $S_0 = 0$ and $S_i = S_i^d$ for $i = 1, \dots, n$.
2. Set as initial solution $S_0 = 1$ and $S_i(1) = S_i^s$ for $i = 1, \dots, n$.
3. Repeatedly, greedily reduce by one the base stock level of LW i with the smallest $F(S_i, a_i)$. Stop when costs no longer decrease. This yields base stock levels $S_i^*(S_0)$.
4. Set $S'_0 = S_0 + 1$ and set $S_i(S'_0) = S_i^*(S_0)$. Carry out Step 3 on $S_i(S'_0)$ to obtain $S_i^*(S'_0)$.
5. Repeat Step 4 until S'_0 equals S_0^u .
6. Return the minimum cost solution.

The optimality of Step 3 follows from Lemma 7. Step 4 is optimal for $R = 0$ and conjectured to be optimal for $R > 0$. The algorithm evaluates costs $S_0^u + 1 + \sum_{i=1}^n (S_i^* + 1)$ times, which is linear in the number of warehouses. Since the approximate costs can be calculated efficiently and the number of steps is linear, the algorithm is efficient.

Note that instead of approximate costs, one can also use exact costs from Section 3.2 in the above procedure. Experiments with randomly generated scenarios indicate that using exact costs in the procedure yields optimal solutions in 99.8% of scenarios. The greedy rule in Step 3 is not optimal under exact costs, however, in our experiments the modified rule

$$i^* = \arg \min_{i \in I} C(S - e_i)$$

yields optimal solutions without exceptions.

8. Experimental results

In this section, we carry out a number of experiments. We examine the performance of the greedy algorithm in Section 8.1. In Section 8.2 we test the validity of assuming exponential rental times, and in Section 8.3 we test the algorithm in case of a service level constraint. In Sections 8.4 and 8.5, we determine the importance of taking into account partial backordering and we evaluate the pooling effect and holding cost benefits of having a low-cost support warehouse.

Table 2 shows the parameters in our experiments. Some of the experiments that are geared mostly to generating insights will use the parameter settings of the base scenario. The other experiments will use 1000 randomly generated instances with parameters uniformly generated between the minimum and maximum values from Table 2. Demand rates have not been included in Table 2 since

Table 2
Experimental design.

Parameter	Base scenario	Minimum	Maximum
h	1	1	1
h_0	0.5	0.2	1
c	5	1	10
b	10	$c + 1$	$c + 10$
ℓ	20	$b + c + 1$	$b + c + 10$
μ	0.5	0.2	1
R	1	0	2

we want these to vary between warehouses, and the number of warehouses typically varies within experiments. Without loss of generality we set the holding cost at the LWs at $h = 1$. To ensure that $\ell \geq b + c$ and $b \geq c$, the minimum value for ℓ is $b + c + 1$ and the minimum value for b is $c + 1$. The partial backorder level R takes on values 0, 1, and 2.

For each value of R the same set of scenarios is used, in order to compare situations with and without partial backordering. Based on a data set with library transactions, the weekly demand rate per LW typically ranges between 0.05 and 0.5. Therefore we let $\lambda_i, i = 1, \dots, n$ be uniform $U(l, l + r)$ distributed, where l is generated uniformly between 0.05 and 0.3 and r between 0 and 0.2. This modeling choice yields a varied range of demand parameters throughout the scenarios, including situations where all LWs have almost the same demand rate in case r is small.

8.1. Performance of the greedy algorithm

Now we evaluate the performance of the greedy algorithm in an experiment with random scenarios. For each randomly generated scenario, we determine optimal base stock policies using complete enumeration and heuristic policies using the greedy algorithm with the approximated cost function.

For different values of n , the performance is summarized in Table 3. The first column shows the percentage of scenarios in which the greedy algorithm leads to an optimal policy. The remaining columns show, for each partial backorder level R , the average, conditional, and maximum percentage deviations compared to the optimal costs. The conditional deviation is the average deviation for scenarios where no optimal policy is found. The deviation for a policy S is defined as $100\% \times (C(S)/C(S^*) - 1)$, with S^* the optimal policy.

In over 80% of the scenarios the greedy algorithm leads to an optimal solution. As n increases, an optimal solution is found less often, however, the average deviations appear to improve for increasing n . The average deviations are below 0.2% for any combination of n and R . The conditional deviations are typically below 1% and the maximum deviation is at most 5.94%. All types of deviations tend to decrease in n . The algorithm thus appears to provide more robust solutions as n increases. Overall, the algorithm seems to have similar performance in situations with and without partial backordering.

8.2. Insensitivity to rental time distributions

We develop a simulation model to determine costs of base stock policies under alternative rental time distributions having the same mean μ^{-1} as the exponential distribution. The two considered distributions are the degenerate distribution at μ^{-1} and uniform $U(0, 2\mu^{-1})$; other less extreme distributions give similar results. By complete enumeration base stock policies are determined with minimal simulated costs. We carry this out for $n = 3$ LWs, using our experimental design from Table 2. Each policy S is simulated 500 times for 1000 time units, excluding a warmup period of 200 time units starting at initial inventory S .

Table 3
Average percentage cost deviations split up by backorder level R .

n	Optimum found (%)	Average deviation (%)			Conditional deviation (%)			Maximum deviation (%)		
		$R=0$	$R=1$	$R=2$	$R=0$	$R=1$	$R=2$	$R=0$	$R=1$	$R=2$
2	88.03	0.20	0.15	0.10	1.40	1.24	1.15	5.94	4.89	5.12
3	87.10	0.15	0.10	0.10	0.96	0.81	0.85	4.05	3.87	3.89
4	86.43	0.12	0.07	0.08	0.75	0.65	0.65	3.69	3.45	3.47
5	84.33	0.11	0.06	0.07	0.56	0.44	0.50	1.98	1.87	3.77
6	82.60	0.10	0.07	0.06	0.49	0.42	0.40	2.73	2.61	1.88

Table 4
Comparison of base stock policies under exponential distributions and alternative distributions.

R	Degenerate				Uniform			
	Same policy (%)	Avg. dev. (%)	Cond. dev. (%)	Max. dev. (%)	Same policy (%)	Avg. dev. (%)	Cond. dev. (%)	Max. dev. (%)
0	95.8	0.00	0.19	1.44	96.9	0.01	0.20	0.97
1	86.4	0.08	0.29	2.69	94.4	0.03	0.21	2.23
2	88.6	0.08	0.28	4.37	93.0	0.02	0.24	4.30

We compare the cost-minimizing policies with those found under exponential rental times in Table 4. In the vast majority of scenarios the policies are the same and the average difference in costs is 0.08% or less. Similar to Alfredsson and Verrijdt (1999), we conclude optimal base stock policies are rather insensitive to the choice of rental time distribution, even under partial backordering.

8.3. Incorporating a service level constraint

Since some applications have a service level constraint instead of pure costs, we now incorporate such a constraint in the model. We set $b = \ell = 0$ and impose the constraint that 95% of the total demand in the system should be met immediately, either from on-hand stock at the LW or by shipments from the SW. Let $TS(S)$ denote this total service level for a base stock policy S . As obtaining near-optimal base stock policies is also important with a service level constraint, we test the effectiveness of the greedy algorithm with the approximated cost function as a heuristic in this new setting.

For obtaining heuristic base stock policies, the upper bounds and greedy algorithm need to be adapted as follows. Upper bounds S_i^s , S_i^d and S_0^u are obtained analogous to the pure cost model. For example, S_i^d should be such that LW i serves at least 95% of its demand from on-hand stock. The greedy algorithm requires changes to the following steps. In Step 3 (Section 7), we continue iterating until $TS(S) < 95\%$, instead of until costs no longer decrease. We determine $TS(S)$ by simulation, using the simulation model from Section 8.2. In Step 1, we also improve the decoupled solution, since achieving at least 95% overall service generally does not require each LW to achieve 95% service individually. The other steps remain the same.

In small instances, we compare these heuristic policies with optimal base stock policies obtained by complete enumeration. In the complete enumeration, we evaluate exact costs and service of base stock policies using the Markov chain from Section 3. For completeness, the exact service is calculated as

$$TS(S) = \frac{1}{\sum_{i=1}^n \lambda_i} \sum_x \pi(x) \sum_{i=1}^n \lambda_i (\mathbb{1}\{x_i > 0\} + \mathbb{1}\{x_i = 0, x_0 > 0\}).$$

Table 5 shows results for 1000 random instances with $n = 3$. The heuristic policies are the same as the optimal policy in over 95% of all scenarios. Due to randomness in the simulation, in approximately 0.1% of the scenarios heuristic policies are obtained

Table 5
Comparing the heuristic policy with the optimal policy under a service level constraint.

R	Optimum found (%)	$TS(S) < 95\%$ (%)	Average deviation (%)	Conditional deviation (%)	Maximum deviation (%)
0	97.9	0.09	0.01	1.14	7.80
1	95.3	0.13	0.05	1.42	13.10
2	95.3	0.13	0.05	1.43	10.35

with service levels between 94.9% and 95%. Such negligible violations of the service level constraint should pose no practical problems. We excluded these policies (with negative deviations) in the calculations of the average deviations. The average deviations are of similar order as the pure cost model, showing that the heuristic has satisfactory performance under a service level constraint.

8.4. Insights for partial backordering

Now we evaluate how optimal base stock policies for the no backordering case perform in case of partial backordering and by how much allowing a partial backorder level reduces lost demand.

Fig. 4 shows the average deviations as a function of n when optimal base stock policies for no backordering are used in case there is partial backordering. We observed that the no backordering policy is optimal in approximately 65% of all scenarios, however, when it is not optimal, the errors can be significant. On average the deviations for $R = 1$ and $R = 2$ are 0.67% and 1.32% and the maximum errors are 23.52% and 33.64%. For any n , the deviation for $R = 2$ is approximately twice as high as for $R = 1$. We observe that the average deviations become smaller as n increases. Hence, it is especially important to account for partial backordering when the number of LWs is small. However, the gap with the optimal solution remains significant, especially for $R = 2$. Overall, the greedy algorithm seems to outperform the no backordering policy.

Fig. 5 shows the reduction in the number of lost demands compared to the $R = 0$ case for different values of R and n . Even for $R = 1$, partial backordering prevents more than 50% of the lost demands in this setting with low demand. As n increases, we can prevent relatively more lost demand by introducing partial backordering. Since a SW is typically used for low demand items, partial backordering can thus be an effective means to reduce lost demand.

Fig. 4. Costs of using the no backordering optimal base stock policy in case of partial backordering for different values of R and n .

Fig. 5. Reduction in the number of lost demands compared to the no backordering case.

We also tested an alternative backordering policy with a central backorder limit nR at the SW instead of R at each LW, similar to Axsäter et al. (2013). In over 70% of scenarios this leads to an average increase in cost of approximately 1–2% with maxima up to 26%, mostly since shipment costs are incurred for each backorder. When the alternative is better, it is typically less than 1% better and at most 5%. This policy could be of interest when a central backorder limit is easier to coordinate.

8.5. Cost reduction from low holding costs at the SW

Now we consider the effect of introducing an SW with shipments and low holding costs in a rental system. We compare the costs of this system to the costs of a decoupled system. We calculate the costs and base stock policies using the greedy algorithm, since the experiments from Section 8.1 indicate this has satisfactory performance. In addition this allows us to consider instances with a large number of LWs.

In the experiment we consider the base case from Table 2 for $n = 1, \dots, 100$ identical LWs with demand rates $\lambda_i \equiv \lambda$. We let λ take on values 0.05, 0.2, 0.5, 2.0 so that the experiment includes low, medium and high demand situations. To observe the effect of the holding cost, we let the holding cost at the SW take on values $h_0 = 0, 0.5, 1$. Fig. 6 depicts the percentage cost reductions of a system with a support warehouse compared to a decoupled system.

From Fig. 6 we can observe that the cost reduction from having an SW increases with n for each value of λ and $h_0 > 0$. This is part due to the well-known stock pooling effect, and part due to the lower holding costs at the SW. With $h_0 = 1$, the holding costs at the SW and LW are equal, so this measures the pure stock pooling effect. By comparing the case $h_0 = 1$ with $h_0 = 0.5$, we can see the additional cost reduction from having lower holding costs at the

SW. This difference is considerable and seems to be largest for low values of λ and n . With $h_0 = 0$, we can store an unlimited number of items at the SW and prevent all backorder and lost demand costs. The gain compared to a decoupled system becomes constant in n because the costs per LW is constant. The $h_0 = 0$ case shows the maximum possible benefit of having an SW, which is largest for $\lambda = 0.05$ and mostly decreasing in λ , albeit not everywhere. For large n , the pooling effect is most dominant.

9. Conclusion

In this paper, we optimize base stock levels for rental systems with expedited shipments from a low-cost support warehouse and partial backordering. For a serial system with one support warehouse and one local warehouse, we fully characterize the costs and optimal base stock policy. For systems with a small number of local warehouses, we can evaluate the exact steady-state costs of a given base stock policy and obtain the optimal policy by enumeration. For large instances, we approximate the costs and formulate a greedy algorithm with near-optimal performance.

Existing methods that were developed for the no backordering case seem to be quite ineffective in situations with partial backordering, especially in smaller systems, showing the necessity of accounting for partial backordering in rental systems. By allowing partial backordering, the fraction of demand that is lost decreases substantially. Even if the partial backorder level is small, a large portion of the benefits of full backordering can be captured by allowing partial backordering. Another interesting insight is that lower holding costs at the support warehouse impose a threshold structure on optimal base stock policies.

Several research avenues can be considered for future work. The approach using results from serial systems likely applies to

Fig. 6. Percentage cost reduction for using a system with shipments from a support warehouse instead of a decoupled system. The scenarios for holding costs and number of warehouses are grouped by demand rate.

other settings, for example to models with unidirectional lateral transshipments. For these models we may obtain bounds that can improve current heuristics and optimization algorithms. One can also evaluate the influence of partial backordering in other settings, for example in models with lateral transshipments. Another interesting direction is to allow lateral transshipments between local warehouses in addition to expedited shipments from a support warehouse. For this new situation we need to determine optimal base stock levels as well as optimal transshipment decisions.

Acknowledgments

This research has been funded by Dinalog, the Dutch Institute for Advanced Logistics. We kindly thank the three anonymous referees for their helpful suggestions to improve the paper.

Appendix

Proof of Theorem 1:

Proof. For the shipment costs $\bar{c}(S_0, S_1)$, use that Eq. (5) can be written as

$$\begin{aligned}\bar{c}(S_0, S_1) &= c\lambda_1 P(x_0 > 0, x_1 = 0) + c\mu S_0 P(x_0 = 0, -R \leq x_1 < 0) \\ &= c\lambda_1 (P(x_1 = 0) - P(x_0 = 0, x_1 = 0)) + c\mu S_0 P_R.\end{aligned}$$

The expression then follows because, by Lemma 1(ii), $P(x_1 = 0) = (1 - P_R)B(S_1, a_1)$, and because

$$\begin{aligned}P(x_0 = 0, x_1 = 0) &= P(x_0 = 0, -R \leq x_1 \leq 0) - P(x_0 = 0, -R \leq x_1 < 0), \\ &= B^R(\bar{S}, a_1) + W^R(\bar{S}, a_1) - P_R.\end{aligned}$$

The expression for $\bar{h}_1(S_0, S_1)$ follows since by Lemma 1(ii) and Eq. (6), the steady state number of available servers at the LW is $(1 - P_R)(S_1 - A(S_1, a_1))$. If we rewrite this as

$$\bar{h}_1(S_0, S_1) = h(S_1 - (1 - P_R)A(S_1, a_1) - P_R S_1),$$

then we can see that carried load of the LW is

$$(1 - P_R)A(S_1, a_1) + P_R S_1.$$

As the total carried load of the system is $A^R(\bar{S}, a_1)$, the carried load of the SW is

$$A^R(\bar{S}, a_1) - (1 - P_R)A(S_1, a_1) - P_R S_1,$$

and the expression for $\bar{h}_0(S_0, S_1)$ follows. Since arriving customers are blocked and delayed with probabilities $B^R(\bar{S}, a_1)$ and $W^R(\bar{S}, a_1)$, the expressions for $\bar{c}_1(S_0, S_1)$ and $\bar{b}_1(S_0, S_1)$ are immediate. $\square$

Proof of Lemma 7:

Proof. Select an arbitrary $i \in I$ and change to base stock policy $S - e_i$. The new demand rate for shipments from the SW is then $\lambda_0 + \delta_i$ with

$$\delta_i = \lambda_i (B(S_i - 1, a_i) - B(S_i, a_i)) = \mu F(S_i, a_i) > 0.$$

Now calculate the cost difference $\tilde{C}(S) - \tilde{C}(S - e_i)$ using (13). Letting $D^{nR}(S_0, a_0) = B^{nR}(S_0, a_0) + W^{nR}(S_0, a_0)$, we obtain after rearranging terms

$$\begin{aligned}\tilde{C}(S) - \tilde{C}(S - e_i) &= h - (h - h_0) \frac{\delta_i}{\mu} - c\delta_i \\ &\quad - \left(\ell - b - c + \frac{h_0}{\mu} \right) \left((\lambda_0 + \delta_i) B^{nR} \left(S_0, a_0 + \frac{\delta_i}{\mu} \right) - \lambda_0 B^{nR}(S_0, a_0) \right) \\ &\quad - b \left((\lambda_0 + \delta_i) D^{nR} \left(S_0, a_0 + \frac{\delta_i}{\mu} \right) - \lambda_0 D^{nR}(S_0, a_0) \right)\end{aligned}$$

All terms involving δ_i have negative coefficients. The probabilities $B^{nR} \left(S_0, a_0 + \frac{\delta_i}{\mu} \right)$ and $D^{nR} \left(S_0, a_0 + \frac{\delta_i}{\mu} \right)$ increase in δ_i since more customers are blocked and delayed with a higher arrival rate. Hence, $\tilde{C}(S) - \tilde{C}(S - e_i)$ decreases in δ_i . We maximize the cost difference by selecting LW i with the smallest δ_i , or equivalently $F(S_i, a_i)$ since μ is constant. $\square$

Proof of Lemma 8:

Proof. We prove this by showing that $\tilde{C}(S) - \tilde{C}(S - e_i)$ is an increasing function of S_0 . For $R = 0$ we get

$$\begin{aligned}\tilde{C}(S) - \tilde{C}(S - e_i) &= h - (h - h_0) \frac{\delta_i}{\mu} - c\delta_i \\ &\quad - \left(\ell - c + \frac{h_0}{\mu} \right) \left((\lambda_0 + \delta_i) B \left(S_0, a_0 + \frac{\delta_i}{\mu} \right) - \lambda_0 B(S_0, a_0) \right)\end{aligned}$$

This function depends on S_0 only through

$$(\lambda_0 + \delta_i) B \left(S_0, a_0 + \frac{\delta_i}{\mu} \right) - \lambda_0 B(S_0, a_0), \quad (14)$$

which is strictly positive since $\lambda_0 B(S_0, a_0)$ is convex and increasing in λ_0 for a fixed μ .

For fixed λ_0 , (14) equals δ_i for $S_0 = 0$ and converges to 0 as $S_0 \rightarrow \infty$. Moreover, (14) decreases in S_0 because $\lambda_0 B(S_0, a_0)$ is convex and decreasing in S_0 . Therefore, we can conclude that $\tilde{C}(S) - \tilde{C}(S - e_i)$ increases in S_0 , implying that the cost reduction for removing an item from an LW becomes larger as S_0 increases. Hence, if it is cost-efficient to remove an item from LW i at base stock level S_0 , then it is certainly cost-efficient at base stock level $S_0 + 1$. It follows that $S_i^*(S_0 + 1) \leq S_i^*(S_0)$. $\square$

References

- Alfredsson, P., & Verrijdt, J. (1999). Modeling emergency supply flexibility in a two-echelon inventory system. *Management Science*, 45(10), 1416–1431.
- Axsäter, S. (1990). Modelling emergency lateral transshipments in inventory systems. *Management Science*, 36(11), 1329–1338.
- Axsäter, S. (2003). Evaluation of unidirectional lateral transshipments and substitutions in inventory systems. *European Journal of Operational Research*, 149(2), 438–447.
- Axsäter, S., Howard, C., & Marklund, J. (2013). A distribution inventory model with transshipments from a support warehouse. *IIE Transactions*, 45(3), 309–322.
- Basten, R., & Van Houtum, G. (2013). Near-optimal heuristics to set base stock levels in a two-echelon distribution network. *International Journal of Production Economics*, 143(2), 546–552.
- Bijvank, M., & Vis, I. (2011). Lost-sales inventory theory: A review. *European Journal of Operational Research*, 215(1), 1–13.
- Bolch, G., Greiner, S., De Meer, H., & Trivedi, K. (2006). *Queueing networks and Markov chains: Modeling and performance evaluation with computer science applications*. John Wiley & Sons.
- Chu, C., Patuwo, B., Mehrez, A., & Rabinowitz, G. (2001). A dynamic two-segment partial backorder control of (r, Q) inventory system. *Computers & Operations Research*, 28(10), 935–953.
- Gross, D., Shortle, J., Thompson, J., & Harris, C. (2008). *Fundamentals of queueing theory*. John Wiley & Sons.
- Howard, C., Marklund, J., Tan, T., & Reijnen, I. (2015). Inventory control in a spare parts distribution system with emergency stocks and pipeline information. *Manufacturing & Service Operations Management*, 17(2), 142–156.
- Kranenburg, A., & Van Houtum, G. (2009). A new partial pooling structure for spare parts networks. *European Journal of Operational Research*, 199(3), 908–921.
- Kukreja, A., Schmidt, C., & Miller, D. (2001). Stocking decisions for low-usage items in a multilocation inventory system. *Management Science*, 47(10), 1371–1383.
- Messierli, E. (1972). Proof of a convexity property of the Erlang B formula. *Bell System Technical Journal*, 51(4), 951–953.
- Montgomery, D., Bazaraa, M., & Keswani, A. (1973). Inventory models with a mixture of backorders and lost sales. *Naval Research Logistics Quarterly*, 20(2), 255–263.
- Öner, K., Kiesmüller, G., & Van Houtum, G. (2009). Monotonicity and supermodularity results for the Erlang loss system. *Operations Research Letters*, 37(4), 265–268.
- Paterson, C., Kiesmüller, G., Teunter, R., & Glazebrook, K. (2011). Inventory models with lateral transshipments: A review. *European Journal of Operational Research*, 210(2), 125–136.
- Rabinowitz, G., Mehrez, A., Chu, C., & Patuwo, B. (1995). A partial backorder control for continuous review (r, Q) inventory system with poisson demand and constant lead time. *Computers & Operations Research*, 22(7), 689–700.
- Rong, Y., Atan, Z., & Snyder, L. (2014). Heuristics for base-stock levels in multi-echelon distribution networks with first-come first served policies. Working Paper. Available at SSRN 1475469. <http://onlinelibrary.wiley.com/doi/10.1111/poms.12717/full>.
- Rosales, C., Rao, U., & Rogers, D. (2013). Retailer transshipment versus central depot allocation for supply network design. *Decision Sciences*, 44(2), 329–356.
- Sherbrooke, C. (1968). METRIC: A multi-echelon technique for recoverable item control. *Operations Research*, 16(1), 122–141.
- Thangam, A., & Uthayakumar, R. (2008). A two-level supply chain with partial back-ordering and approximated Poisson demand. *European Journal of Operational Research*, 187(1), 228–242.
- Van der Heide, G., & Roodbergen, K. (2013). Transshipment and rebalancing policies for library books. *European Journal of Operational Research*, 228(2), 447–456.
- Van Wijk, A., Adan, I., & Van Houtum, G. (2013). Optimal allocation policy for a multi-location inventory system with a quick response warehouse. *Operations Research Letters*, 41(3), 305–310.
- Wee, K., & Dada, M. (2005). Optimal policies for transshipping inventory in a retail network. *Management Science*, 51(10), 1519–1533.