Int. J. Production Economics 143 (2013) 518–526

Contents lists available at SciVerse ScienceDirect

Int. J. Production Economics

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

The selective use of emergency shipments for service-contract differentiation

E.M. Alvarez *, M.C. van der Heijden, W.H.M. Zijm

School of Management and Governance, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Available online 15 March 2012

Keywords:
Inventory
Customer differentiation
Emergency shipments
Service contracts
Critical levels

ABSTRACT

Suppliers of capital goods increasingly offer performance-based service contracts with customer-specific service levels. To handle such differentiated service levels, we use selective emergency shipments of spare parts. We apply emergency shipments in out-of-stock situations for combinations of parts and customer classes that yield service levels close to the class-specific targets. We develop two heuristics to solve this problem. An extensive numerical experiment reveals average cost savings of 4.4% compared to the one-size-fits-all approach that is often used in practice. Furthermore, it is particularly beneficial to combine our policy with critical levels, which yields an average cost saving of 13.9%.

1. Introduction

To service advanced capital goods (e.g., defense or medical systems), suppliers increasingly offer service contracts to their customers. This particularly applies when system downtime can have serious consequences (e.g., loss of production output, failure of military missions). Service contracts typically contain quantified targets for key performance measures such as a maximum response time in case of failures or a minimum system availability. As users typically value downtime differently, service level agreements may differ among customer groups. For example, the minimum system availability may be 90% or 99%.

In practice, suppliers often service customers that have varying service levels using a uniform logistics fulfillment process (a so-called one-size-fits-all approach, cf. Cohen et al. 2006). This approach can be very costly if a supplier designs the fulfillment process based on the premium service level. Also, standard customers have no incentive to switch to premium contracts. The fulfillment process should thus be such that the actual service levels reflect the contractual agreements. In this paper, we focus on differentiation in *spare parts supply*.

In literature, *critical level policies* are common differentiation approaches. Such policies reserve parts for premium customers once the inventory level drops below a certain threshold. Then, demand from non-premium customers is either backordered or satisfied from a source that is usually assumed to have infinite supply (e.g., a production facility). Although shown to be effective and efficient, there are barriers for implementation in practice. For instance, service engineers responsible for system repair are often unwilling to wait for a part that is in fact in stock when they are primarily accountable for the speed of repair.

These drawbacks prompt us to investigate the *selective use of emergency shipments* as an alternative. Now, we use on-hand stock to meet demand first-come-first-served. If we are out of stock, we can request an emergency shipment from a secondary source. Emergency shipments are both faster and more expensive than regular replenishments. We should thus investigate for which combinations of customer segments and item types it is a viable approach. As main advantage, this approach is easier to implement in practice than critical level policies, while still being a tool for differentiation. We will show that the approach leads to clear savings over using simple one-size-fits all strategies. Also, we will show that it is very effective to *combine* selective emergency shipments and critical level policies.

The remainder of the paper is structured as follows. In Section 2, we discuss relevant literature and state our contribution. Then, we state our optimization problem and solution approach in Sections 3 and 4, respectively. It will become clear that we must analyze various single-item models as building blocks. In Section 5, we analyze these models for the special case with two customer classes. We give the results of an extensive numerical experiment in Section 6. In Section 7, we give conclusions and discuss options for model extension.

2. Literature overview

Our research is related to literature on service differentiation and the use of emergency shipments for parts supply. The service differentiation stream focuses on critical level policies, introduced by Veinott (1965). The optimality of this policy has been shown under periodic review for backordering and lost sales (Topkis, 1968). Under continuous review, optimality has been shown for

* Corresponding author.
E-mail address: e.m.alvarez@utwente.nl (E.M. Alvarez).

0925-5273/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijpe.2012.02.019

Poisson demand and exponential or Erlang lead times, both for lost sales (Ha, 1997a; 2000) and backorders (Ha, 1997b; De Véricourt et al., 2002; Gayon et al., 2009).

Several approaches have been developed to find (near-) optimal stock levels and critical levels. For *fast movers*, the focus is on continuous demand distributions with unmet demand being backordered. Ha (1997b) shows that it is optimal to only use arriving replenishment orders to clear non-premium backorders if the inventory level is at least the critical level for premium demand. As the mathematical analysis of such models is intractable – we must keep track of the non-premium backorders – heuristics are often used, see e.g., Möllering and Thonemann (2010) and Arslan et al. (2007). Models for *slow movers*, common in service logistics, focus on Poisson demand and one-for-one replenishment (Dekker et al., 2002). Our work is most similar to Kranenburg and Van Houtum (2007), who try to minimize holding and shipment costs in a multi-item multi-class model with class-dependent waiting time restrictions. Unmet demand is satisfied through emergency shipments. The authors use a solution approach based on decomposition and column generation, combined with greedy heuristics.

The second relevant literature stream focuses on emergency shipments, possibly combined with lateral transshipments among locations at the same echelon level. Most papers consider a single- or two-echelon model with a central location that has infinite supply (see e.g., Muckstadt and Thomas, 1980; Hausman and Erkip, 1994). Alfredsson and Verrijdt (1999) combine lateral shipments with emergency shipments for a two-echelon single-item network. Other recent contributions include Van Utterbeeck et al. (2009) and Wong et al. (2007). We have not yet found literature that uses emergency shipments as a differentiation tool.

The contribution of our paper is fourfold: First, we give a *new differentiation approach* in spare parts supply using selective emergency shipments. Second, we develop *two efficient and effective heuristics* to find near-optimal stock levels and shipment strategies. Third, we show the *added value of selective emergency shipments* compared to one-size-fits-all policies and critical level policies. Finally, we show the *added value of combining selective emergency shipments and critical level policies* for service differentiation.

3. Model

We first give an outline of our model. Next, we discuss the validity of our selection of shipment policies (Section 3.2). In Section 3.3, we present our model assumptions and notation. We give the formal optimization problem in Section 3.4.

3.1. Model outline

Consider a local warehouse that supplies various types of parts to multiple customer classes, and a central depot with infinite supply that replenishes the local warehouse. All customers have the same system, with each item in the system being critical (i.e., an item failure causes a system failure). Each customer class has a distinct amount of time it is willing to wait for parts on average. The warehouse fills demand from all classes first-come-first-served. If it is out of stock, the warehouse may backorder the demand or request an emergency shipment from the central depot. We achieve service differentiation by only using emergency shipments for customer classes with tight waiting time restrictions. We expect this to be particularly beneficial for expensive slow movers that often have low fill rates (making the difference between regular and emergency shipment times crucial). Still, it will sometimes be better to avoid stocks altogether and use emergency shipments for all classes. Conversely, for cheap fast movers it is probably better to keep large stocks (avoiding expensive emergency shipments) and use full backordering. The shipment mode should thus depend on both the item characteristics and waiting time constraints per customer class.

In addition to the above model, we also consider a model where critical levels and selective emergency shipments are jointly used for differentiation. This combined model only satisfies demand from on-hand stock if it exceeds the critical level for the customer's class. Unmet demand is met using either backordering or emergency shipments.

The objective in both models is to minimize system holding and shipment costs, under restrictions on the mean aggregate waiting time per class. Firms like Philips Healthcare and Océ Technologies usually have service level requirements with their clients in terms of e.g., average failure resolution times, with delays often being caused by waiting time for spares. Penalties may apply if the supplier violates the agreements, but we have not seen explicit backorder costs in service contracts. Therefore, we do not include penalty costs per unit waiting time in our objective function. Our decision variables are the item stock levels, and the shipment mode (regular, emergency) and critical level for each item and customer class.

3.2. Selection of shipment policies

In our model, we always use emergency shipments in out-of-stock settings if that shipment mode is chosen for a customer class. However, if the pipeline contains many items, the emergency shipment time might exceed the backorder waiting time, making backordering the faster *and* cheaper option. Ideally, we should thus consider the system state and the customer's class when deciding what shipment mode is most effective. Still, we do not consider such policies to keep the notation transparent and reduce computational effort. In the end, we are mainly interested in the suitability of selective emergency shipments for differentiation compared to critical level policies and the "one-size-fits-all" approach.

3.3. Assumptions and notation

3.3.1. Main assumptions

1. *Demand for each item occurs according to a Poisson process.*
2. *An* $(S-1,S)$ *base stock policy is applied for all items.* In practice, spares often tend to be expensive slow movers. Therefore, holding costs usually dominate ordering costs and hence the optimal ordering quantity is usually 1.
3. *Regular shipment times from depot to warehouse are exponentially distributed.* This assumption facilitates Markov chain analysis. Also, we show in Appendix A that our model is quite insensitive to lead time distribution used.
4. *The shipment time from the local warehouse to the customer is negligible.*
5. *An emergency shipment is shipped directly from central depot to customer* (i.e., *the shipment does not pass through the local warehouse*).
6. *We consider an infinite horizon.* As a result, the mean waiting time for any customer in class j will equal the average waiting time of class j as a whole.

3.3.2. Notation

For each item $i=1,2,\ldots,I$, we denote the mean replenishment lead time by T_i^{reg}, the emergency shipment time by T_i^{em}, the holding costs per time unit by h_i and the additional costs for an emergency shipment over a normal replenishment by EC_i^{em}. The

latter cost factor is sufficient, since each request triggers either a normal replenishment or an emergency shipment. Customers are assigned to classes $j=1, ..., J$, each having a target $W_j^{\max}$ for the average waiting time for parts. We order classes according to non-decreasing values of $W_j^{\max}$. Class j demand for item i occurs at rate m_{ij} (>0). $M_{\cdot j}=\sum_{i=1}^{I} m_{ij}$ and $M_{i\cdot}=\sum_{j=1}^{J} m_{ij}$ denote the total mean demand for class j and for item i, respectively. The decision variables for item i are: (1) the base stock level S_i (2) the vector $\mathbf{C}_i=[C_{i1}, ..., C_{iJ}]$ denoting the critical levels per class, and (3) the *shipment strategy* D_i, denoting the highest customer class index for which emergency shipments are used in a stock-out situation. D_i is an integer between 0 and J, as emergency shipments are only sensible for higher priority customers. We combine all variables into an *item policy* $(S_i,D_i,\mathbf{C}_i)$ with mean waiting time $EW_{ij}(S_i,D_i,\mathbf{C}_i)$ and fill rate $\beta_{ij}(S_i,D_i,\mathbf{C}_i)$ for item i and class j as performance indicators.

3.4. Formal optimization problem

We express the formal optimization problem (*P*1) as follows:

$$\min_{S_i,D_i,\mathbf{C}_i} \sum_{i=1}^{I} \left\{ h_i S_i + EC_i^{\mathrm{em}} \sum_{j=1}^{J} m_{ij}(1-\beta_{ij}(S_i,D_i,\mathbf{C}_i)) I_{\{1\ldots D_i\}}(j) \right\}$$

$$s.t. \quad \sum_{i=1}^{I} \frac{m_{ij}}{M_j} EW_{ij}(S_i,D_i,\mathbf{C}_i) \le W_j^{\max} \quad j=1,\ldots,J \qquad \text{(P1.1)}$$

$$S_i \in N_0, \quad D_i \in \{0,1,\ldots,J\}, \quad C_{ij} \in \{0,1,\ldots,S_i\} \quad i=1,\ldots,I, \quad j=1,\ldots,J \qquad \text{(P1.2)}$$

We minimize holding and emergency shipment costs with the demand-weighted mean waiting time for class j not allowed to exceed target $W_j^{\max}$. The indicator function $I_{\{1\ldots D_i\}}(j)$ equals 1 if emergency shipments are used for class j ($j \in \{1\ldots D_i\}$) and 0 otherwise. We compute holding costs over the total stock S_i, including items in the pipeline. It is easy to compute holding costs over the on-hand stock instead (Kranenburg and Van Houtum (2007)).

4. Solution approach

Problem (*P*1) is a nonlinear integer problem that we cannot decompose into separate single-item problems because of the aggregate waiting time restriction (P1.1). This differs from a variant with backorder costs where such a decomposition is possible and the I single-item problems can be solved easily (see Section 4.2). We use an approach similar to Dantzig–Wolfe decomposition: We reformulate (*P*1) to a *linear* integer programming problem and solve its LP-relaxation to find a lower bound. This approach has also been used to solve other inventory problems (Kranenburg and Van Houtum, 2007; 2008). Therefore, we only specify how the approach can be used for our specific problem and refer the reader to our working paper (Alvarez et al., 2010) for more details. We first show how to reformulate (*P*1) to a linear problem and find a lower bound (Section 4.1 and Section 4.2), respectively. As the lower bound is generally fractional, Section 4.3 gives two heuristics to find near-optimal integer solutions.

4.1. Reformulation to a linear problem

Let b_i be a shorthand notation for item policy $(S_i,D_i,\mathbf{C}_i)$. The binary variable x_{b_i} denotes whether policy b_i is selected for item i or not ($x_{b_i}=1$ or 0). Let B_i be the set of items we consider for item i. We then obtain the linear integer program (*P*2).

$$\min \sum_{i=1}^{I} \sum_{b_i \in B_i} TC_i(b_i) x_{b_i}$$

$$s.t. \sum_{i=1}^{I} \sum_{b_i \in B_i} \frac{m_{ij}}{M_j} EW_{ij}(b_i) x_{b_i} \le W_j^{\max} \quad j=1,\ldots,J \qquad \text{(P2.1)}$$

$$\sum_{b_i \in B_i} x_{b_i} = 1 \quad i=1,\ldots,I \qquad \text{(P2.2)}$$

$$x_{b_i} \in \{0,1\} \quad b_i \in B_i \quad i=1,\ldots,I$$

Here, $TC_i(b_i)$ is shorthand for the total costs related to item i under policy b_i, so $TC_i(b_i)=TC_i(S_i,D_i,\mathbf{C}_i)=h_iS_i+EC_i^{\mathrm{em}}\sum_{j=1}^{J} m_{ij}(1-\beta_{ij}(S_i,D_i,\mathbf{C}_i))I_{\{1\ldots D_i\}}(j)$

4.2. Finding a lower bound for the total costs

We first solve the LP-relaxation of (*P*2) with an initial item policy set that results in a feasible solution. Then, we use column generation to iteratively find new item policies that improve the solution if added. We stop once such policies no longer exist. For each item i, we find a single initial policy by setting D_i to 0 and only increasing S_i. Then, we iteratively add the policy with the lowest reduced costs to B_i, if these are negative. The reduced costs for policy b_i, denoted by $RED(b_i)$, are given by the expression below, with $u_j(\le 0)$ and $v_i(\ge 0)$ denoting the current shadow prices for constraints (*P*2.1) and (*P*2.2), respectively.

$$RED(b_i)=TC_i(b_i)-\sum_{j=1}^{J} u_j \frac{m_{ij}}{M_j} EW_{ij}(b_i)-v_i$$

For each shipment strategy D_i, we first find the values for S_i and $\mathbf{C}_i$ that give minimum reduced costs. Next, we select the item policy with minimum reduced costs over all shipment strategies. A complication in finding an optimal item policy for a shipment strategy is that the reduced costs are rarely convex in S_i and/or $\mathbf{C}_i$: the reduced costs are only convex in S_i if $C_{ij}=0$ for all classes and $D_i=J$ (i.e., emergency shipments for all classes, see Kranenburg and Van Houtum (2007)). However, from some value onwards, $RED(b_i)$ will only monotonically increase in S_i, irrespective of the values for D_i and $\mathbf{C}_i$: as S_i increases, the holding costs increase linearly, whereas the costs related to emergency shipments and waiting time decrease and eventually become negligible. We thus find an upper bound on S_i (and all C_{ij}) when the holding costs outweigh the other cost elements in $RED(b_i)$.

4.3. Methods for finding a near-optimal integer solution

We find near-optimal integer solutions using 2 methods: (1) we use the (non-integer) LP relaxation solution as a starting point for local search; (2) we solve *integer* problem (*P*2) with all policies generated for finding the lower bound. In literature (e.g., Kranenburg and Van Houtum (2008)), method 1 is often used, but we show that method 2 works better (Section 6).

4.3.1. Method 1: Use a local search algorithm

Kranenburg and Van Houtum (2008) find an integer solution by first selecting for each item in the LP-relaxation solution the policy with the lowest stock level, provided that $x_{b_i}>0$. As the resulting solution is usually infeasible, they then iteratively increase the stock level of one item until they find a feasible solution. In every iteration, the item is selected that leads to the largest decrease in the total gap between target and actual mean waiting times per invested euro. Our method differs in two ways.

First, the *neighborhood* (i.e., the solutions close to the current solution from which we choose a new solution) contains solutions with either a larger stock level than the current solution *or* a faster shipment method. We consider multiple decision variables to possibly find a feasible solution more quickly. Second, we use a different *selection criterion*: When computing the gaps between target and actual waiting time, we use the inverse of $W_j^{\max}$ as a *weight* for class j. We expect that it is generally more expensive to reduce a small waiting time by amount Δ compared to reducing a large waiting time by that same amount. As a result, we give higher weights to customer classes with tight restrictions. For details, we refer to our working paper (Alvarez et al., 2010).

4.3.2. Method 2: Use of integer programming (IP)

When solving the LP-relaxation, we usually only generate four to seven item policies per item. Therefore, we should be able to solve the corresponding integer problem with a commercial solver (we used CPLEX) for most problems of realistic size. However, the generated item policies are not always related, especially for fast moving items. For instance, we have found policies (0,1,0) (i.e., no stock, emergency shipments for premium customers only), (9,0,0) and (10,0,0) (i.e., high stock levels, full backordering) for the same item. To test the quality of our method when using the relaxation policy set, we compared the resulting solutions to those found when we include additional item policies in the IP that bridge the gap between the distinct item policies.[1] We found that these additional policies greatly increase computation time, while the solution quality improves only marginally: the average gap to the lower bound drops from 0.041 to 0.038, and the maximum gap drops from 0.259 to 0.258. Therefore, we find that it is sufficient to only use the item policies generated when solving the LP relaxation.

5. Evaluating item policies with two customer classes

We use continuous-time Markov chain analysis to find performance measures for an item policy. For simplicity, we limit ourselves to two customer classes in this section and in Section 6. In Section 7, we discuss extensions to more than two classes. We thus consider three shipment strategies: use emergency shipments for both classes ($D_i=2$), for class 1 only ($D_i=1$), or not at all ($D_i=0$). Per item, we have a critical level for the non-premium class (denoted by C_i from now on). Under backordering, we find the expected waiting time by Little's Law: $EW_{ij}(S_i,D_i,C_i)=EBO_{ij}(S_i,D_i,C_i)/m_{ij}$, with $EBO_{ij}(S_i,D_i,C_i)$ being the average number of backorders for item i and class j. When emergency shipments are used, $EW_{ij}(S_i,D_i,C_i)$ equals $(1-\beta_{ij}(S_i,D_i,C_i))T_i^{\text{em}}$. Section 5.1 first describes the pure selective emergency shipment models, where critical levels are not used. Section 5.2 then describes the combined shipment models, which incorporate critical level policies. For simplicity, we omit the item index i and denote the normal replenishment rate by $\mu=1/T^{\text{reg}}$.

5.1. Pure selective emergency shipment models

5.1.1. Pure model, emergency shipments for both classes ($D_i=2$)

We model the pipeline as an Erlang loss system with S servers (Kranenburg and Van Houtum, 2007). Let k denote the number in the pipeline. We find:

$$p_k=\left(\frac{M}{\mu}\right)^k\frac{1}{k!}\Big/\sum_{n=0}^{S}\left(\frac{M}{\mu}\right)^n\frac{1}{n!}\quad k=0,\ldots,S$$

[1] This was a mid-sized experiment of 100 problem instances with 25, 100 and 400 items.

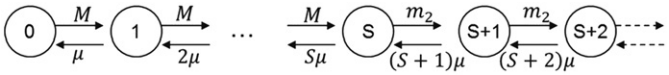

Fig. 5–1. Transition diagram for pure model with backordering of class 2 requests.

$$\beta_1(S,D)=\beta_2(S,D)=1-p_S$$

5.1.2. Pure model, backorder class 2 demand only ($D_i=1$)

We define the state as the number of items in the pipeline k, with state k having $[k-S]^+$ class 2 backorders. Fig. 5–1 displays the Markov chain. Once the pipeline contains S or more items, class 1 demand is lost to the system. Closed-form expressions for the state probabilities p_k, class 1 fill rate and class 2 mean backorder level are given below.

$$p_0=\left\{\sum_{k=0}^{S}\frac{1}{k!}\left(\frac{M}{\mu}\right)^k+\left(\frac{M}{m_2}\right)^S\left(e^{m_2/\mu}-\sum_{k=0}^{S}\frac{1}{k!}\left(\frac{m_2}{\mu}\right)^k\right)\right\}^{-1}$$

$$p_k=\begin{cases}\left(\frac{M}{\mu}\right)^k\frac{1}{k!}p_0 & 1\le k\le S\\ \left(\frac{M}{\mu}\right)^S\left(\frac{m_2}{\mu}\right)^{k-S}\frac{1}{k!}p_0 & k\ge S+1\end{cases}$$

$$\beta_1(S,D)=\sum_{k=0}^{S-1}p_k$$

$$EBO_2(S,D)=\left(\frac{M}{m_2}\right)^S p_0\left\{\frac{m_2}{\mu}\left(e^{m_2/\mu}-\sum_{k=0}^{S-1}\left(\frac{m_2}{\mu}\right)^k\frac{1}{k!}\right)-S\left(e^{m_2/\mu}-\sum_{k=0}^{S}\left(\frac{m_2}{\mu}\right)^k\frac{1}{k!}\right)\right\}$$

5.1.3. Pure model, backorder demand from all classes ($D_i=0$)

We use priority backorder clearing: class 1 backorders are cleared before class 2 backorders, even if a class 2 backorder occurred first. Therefore, we need a two-dimensional state space, since the number of backorders per class can even differ among states with the same number of items in the pipeline. We use states (k,l), with k the number in the pipeline and l the number of class 2 backorders. We then have $[[k-S]^+-l]^+$ class 1 backorders. Fig. 5-2 shows the corresponding Markov chain. Demand flows from $(k,0)$ to $(k+1,0)$ until the pipeline contains S items. Then, class 1 demands result in shifts from (k,l) to $(k+1,l)$, while class 2 demands result in shifts from (k,l) to $(k+1,l+1)$. Replenishment flows go from (k,l) to $(k-1,l)$ whenever (k,l) has class 1 backorders. Then, the pipeline decreases, while the number of class 2 backorders remains the same. Flows from (k,l) to $(k-1,l-1)$ only occur when (k,l) only has class 2 backorders.

We could not find analytical expressions for the state probabilities, so we compute an upper bound k^{UB} on the number of items in the pipeline and solve the balance equations numerically. We find k^{UB} by aggregating all demand into a single class and analyzing the resulting $M|M|\infty$ model exactly. Note that the pipeline distribution will be the same as in the two-class model, since the demand and replenishment rates are the same. We find k^{UB} such that $1-\sum_0^{k^{UB}}p_k\le\varepsilon$ with $\varepsilon=10^{-8}$. We then find the following expressions for $EBO_j(S,D)$:

$$EBO_1(S,D)=\sum_{k=S+1}^{k^{UB}}\sum_{l=0}^{k-S}(k-S-l)p_{kl}$$

$$EBO_2(S,D)=\sum_{k=S+1}^{k^{UB}}\sum_{l=0}^{k-S}l\cdot p_{kl}$$

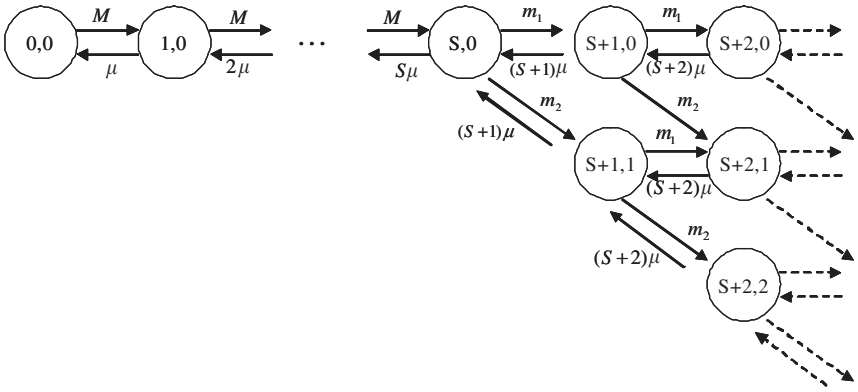

Fig. 5–2. Transition diagram for pure model with full backordering.

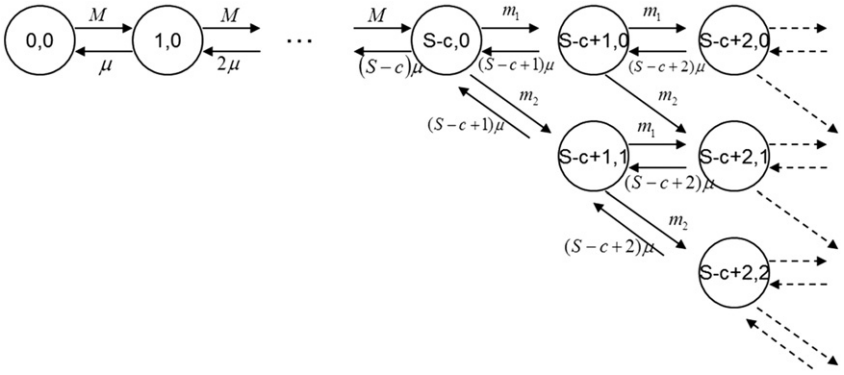

Fig. 5–3. Transition diagram for combined model with full backordering.

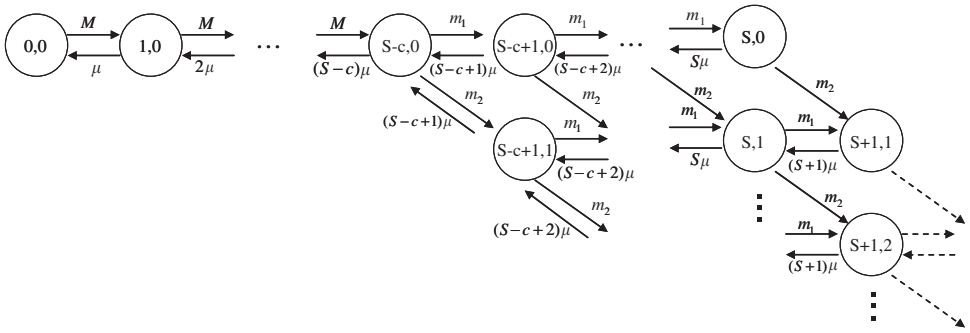

Fig. 5–4. Transition diagram for combined model with backordering of class 2 requests.

5.2. Combined models that incorporate critical levels

We refer to Kranenburg and Van Houtum (2008) for the model with emergency shipments for both classes ($D_i=2$) and only discuss the (partial) backordering models.

5.2.1. Combined model, backorder demands from all classes ($D_i=0$)

We follow Ha (1997b), who shows the optimality of only clearing class 2 backorders once all class 1 backorders have been cleared and the on-hand stock is at least C. Our Markov chain (Fig. 5-3) consists of states (k,l), with k the number in the pipeline and l the number of class 2 backorders: We can simultaneously have stock on-hand and class 2 backorders, so we cannot derive the number of backorders from the pipeline only. The Markov chain branches out at $S–C$ or more items in the pipeline (and thus at most C items on-hand): then, class 1 demand is met from stock, with class 2 demand being backordered. Once we are out of stock, we also backorder class 1 demand. Most replenishment flows go from (k,l) to $(k–1,l)$: we then clear class 1 backorders or increase on-hand stock.

Since this model is very similar to the pure model with full backordering, we use the same approach to compute k^{UB} and find the state probabilities. We find for $EBO_j(S,D,C)$:

$$EBO_1(S,D,C)=\sum_{k=S+1}^{k^{UB}}\sum_{l=0}^{k-S+C}\max\{0,k-S-l\}p_{kl}$$

$$EBO_2(S,D,C)=\sum_{k=S-C+1}^{k^{UB}}\sum_{l=0}^{k-S+C}l\cdot p_{kl}$$

This Markov chain (and hence the state probabilities) only depends on the value of $S–C$. Hence, we directly find the

performance measures for all other policies with the same value for S–C, which greatly reduces the computational burden of analyzing item policies.

5.2.2. Combined model, backorder class 2 demand only ($D_i=1$)

We use state space (k,l) as before, see Fig. 5-4. However, once we are out of stock now, denoted by states $(k,k–S)$ with $k \geq S$, the pipeline only increases further from class 2 demand. Increasing on-hand stock to C has priority over clearing class 2 backorders.

We determine k^{UB} by aggregating all demand and assuming full backordering. Using this value, we then solve the resulting balance equations. Note that our pipeline bound might be larger than necessary, since we assume full backordering when computing k^{UB}. For the performance measures we find:

$$EBO_2(S,D,C) = \sum_{k=S-C+1}^{k^{UB}} \sum_{l=\max\{0,k-S\}}^{k-S+C} l \cdot p_{kl}$$

$$\beta_1(S,D,C) = 1 - \sum_{k=S}^{k^{UB}} p_{k,k-S}$$

6. Computational experiment

We conducted a numerical experiment, for which we state the objectives in Section 6.1. Section 6.2 covers the problem instances and Section 6.3 the results.

6.1. Objectives

Our objectives are: (i) to evaluate the two heuristics for obtaining a near-optimal solution (i.e., local search and IP) in terms of solution quality and computation time, (ii) to determine whether and when selective emergency shipments are effective for service differentiation, (iii) to compare selective emergency shipments to critical level policies as differentiation tools, (iv) to determine the added value of *jointly* using selective emergency shipments and critical level policies for differentiation.

Table 6–1
Parameter values of the tested instances.

Parameter	Values
1 Number of items I	25,100,400
2 Daily demand rate per item $M_{i\bullet}$	U[0,0.1], U[0,0.5]
3 Fractions of class demand per item $(m_{i1}/M_{i\bullet};m_{i2}/M_{i\bullet})$	(0.2;0.8),(0.5;0.5),(0.8;0.2)
4 $(T_i^{reg};T_i^{em})$ (in days)	(4;1),(8;1),(8;2),(16;2)
5 Item holding cost interval (per unit per day)	[0.02;19.98],[0.2;199.8],[2;1998]
6 Target service levels $(W_1^{max};W_2^{max})$ (in hours)	(0.5,2),(0.5,4),(3,12),(3,24)

6.2. Experiment design

Table 6-1 shows the tested parameter values, based on the values by Kranenburg and Van Houtum (2008) which in turn are derived from observations in practice. We use $EC_i^{em} = 1000$ as cost normalization. For each combination of parameters 1, 3, 4 and 6, we generate 4 random instances as follows: for each item, a demand rate and holding cost is drawn from uniform distributions on the given intervals, with the correlation between demand rates and holding costs being -0.8. This is realistic, since fast movers tend to have low (holding) costs in practice and vice versa. Except for the demand rates and holding costs, all items in an instance have the same parameter values. We have 3456 instances in total: we have 864 parameter combinations and 4 demand rate/holding cost samples per combination.

6.3. Results

First, we evaluate the performance of the two heuristics described in Section 4.3. Next, we investigate whether and when the emergency shipment strategy has added value over one-size-fits-all strategies. Finally, we compare the emergency shipment policy to the critical level policy and determine the added value of combining both policies.

6.3.1. Performance of the heuristics

We express the solution quality in terms of a relative gap to the lower bound, defined as $TC_H - TC_{LB}/TC_{LB}$, where TC_H gives the solution value of the heuristic (IP or Local Search). Table 6-2 shows the solution quality and computation times for different numbers of items, the parameter with most impact. We used a Intel quad core, 2.83 GHz processor. We see that integer programming yields a gap to the lower bound less than half that of local search on average. This gap clearly decreases with the number of items. This is beneficial, because practical instances typically contain hundreds of items. The computation times for both methods are small, although the run times of IP increase greatly with the problem size.

Table 6-3
Savings of selective emergency shipments and OSFA BO+ES over OSFA ES.

Parameter	**Values**	**Savings over OSFA ES (%)**			
		OSFA BO+ES		Selective em. shipments	
		Average	Maximum	Average	Maximum
Holding cost range interval	[0.02 – 19.98]	7.9	38.9	11.7	45.8
	[0.2 – 199.8]	0.2	2.7	1.2	12.6
	[2 – 1998]	0.1	1.9	0.3	3.0
Overall		**2.7**	**38.9**	**4.4**	**45.8**

Table 6-2
Solution quality and computation times integer programming (IP) and local search (LS).

Parameter	**Values**	**Gap IP (%)**		**Gap LS (%)**		**Computation time IP (s)**		**Computation time LS (s)**	
		Average	Maximum	Average	Maximum	Average	Maximum	Average	Maximum
Number of SKUs	25	0.25	2.16	0.51	3.81	0.10	0.64	0.01	0.06
	100	0.02	0.11	0.05	0.35	0.14	0.73	0.03	0.73
	400	0.00	0.02	0.00	0.08	1.48	10.34	0.50	2.66
Overall		**0.09**	**2.16**	**0.19**	**3.81**	**0.58**	**10.34**	**0.18**	**2.66**

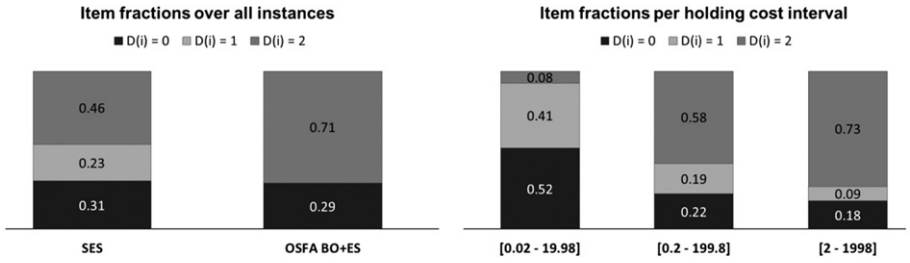

Fig. 6–1. Fraction of items per shipment strategy for selective emergency shipments (SES) and OSFA BO+ES (left) and fraction of items per holding cost interval for SES (right).

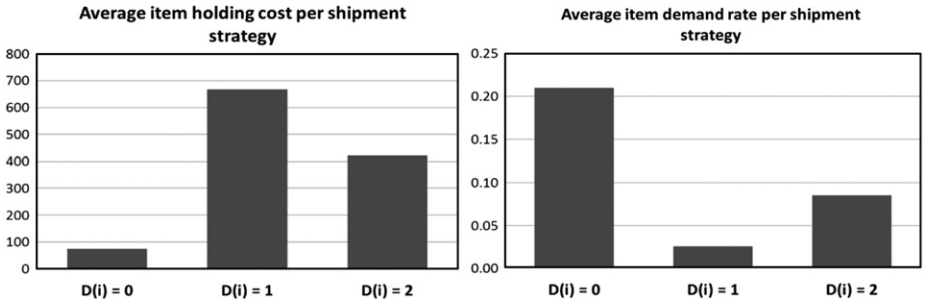

Fig. 6–2. Overall average holding costs and demand rates per shipment strategy.

6.3.2. *The added value of using emergency shipments*

We compare our selective emergency shipment policy to a one-size-fits-all policy with emergency shipments for all items (OSFA ES), and a variant where we use either backordering or emergency shipments for an item (OSFA BO+ES). The latter policy differentiates among items, but *not* among customer classes. Backorder clearing is done first-come-first-served. Under one-size-fits-all, we use a single customer class with maximum waiting time $W_1{}^{\max}$. We use OSFA ES as a benchmark, since this is common for testing critical level policies (e.g., Kranenburg and Van Houtum (2008)). Table 6-3 shows the savings of OSFA BO+ES and our policy compared to OSFA ES. We also show the results for different holding cost intervals, because this parameter has most influence on the savings.

OSFA BO+ES gives average savings of 2.7% over OSFA ES. Selective emergency shipments yield additional average savings of 1.7%, resulting in average savings of 4.4% over OSFA ES with a maximum of 45.8%. The savings are largest when holding costs are low, because emergency shipments are less appealing then: it is cheaper to keep large stocks and reserve emergency shipments for premium customer demand only. We also find large savings when waiting time restrictions for class 1 demand are loose (7.4% on average).

Fig. 6-1 displays on the left the fraction of items assigned to each shipment strategy for OSFA BO+ES and our policy (SES). The figure to the right shows the division of items over shipment strategies per holding cost interval for our policy. Both SES and OSFA BO+ES have roughly the same fraction of items where full backordering is used ($D_i=0$). Clearly, we use the selective shipment strategy ($D_i=1$) to limit using expensive emergency shipments for both classes ($D_i=2$). On average, $D_i=1$ for 20% of the items. This fraction increases to more than 40% when the holding costs are low (see figure on the right).

Fig. 6-2 shows the average item holding costs and demand rates per shipment strategy. Clearly, the selective strategy is mainly used for expensive slow movers. Little or no stock is kept of such items, making the shipment mode used crucial for meeting waiting times: emergency shipments are used for premium clients and non-premium requests are backordered.

Table 6-4
Savings of different policies over OSFA ES.

Parameter	Values	Savings over OSFA ES (%)		
		Sel. em. ship. (SES)	CLP ES	CLP+SES
Treg – Tem (days)	4 – 1	7.6	5.8	16.1
	8 – 1	4.5	8.5	14.3
	8 – 2	4.1	7.5	13.1
	16 – 2	1.5	9.7	12.1
Holding cost range interval	[0.02 – 19.98]	11.7	0.5	16.2
	[0.2 – 199.8]	1.2	8.3	10.5
	[2 – 1998]	0.3	14.8	15.0
W1-max – W2-max (hours)	0.5 – 2	1.4	4.4	6.2
	0.5 – 4	1.4	7.6	10.4
	3 – 12	7.4	8.4	16.7
	3 – 24	7.5	11.1	22.4
Overall		**4.4**	**7.9**	**13.9**

6.3.3. *Comparison to the critical level policy and a combined policy*

We compare our policy to a critical level policy with emergency shipments (CLP ES). Also, we investigate the benefits of combining both policies (CLP+SES). Table 6-4 shows the relative savings of the policies compared to OSFA ES. Critical level policies generally outperform selective emergency shipments, with average savings of 7.9%. This is caused by the mode of differentiation: in critical level policies, premium customers will often obtain a part right away. In contrast, customers need to wait at least T_i^{em} time units for an emergency shipment. The selective emergency shipment policy is also less sensitive to the waiting time restrictions than the critical level policy: the waiting time restriction for class 1 is usually dominant. Then, increasing $W_2^{\max}$ has little impact on the solutions found.

Selective emergency shipments outperform critical level policies in cases with short regular shipment times, low holding costs, and loose waiting time restrictions for class 1 demand. Then, it is viable to meet (a part of) the demand through the regular channel instead of expensive emergency shipments. Indeed, the fraction of items for which D_i is 0 or 1 is relatively high then (for the holding

costs we can see this in Fig. 6-1). Note that selective emergency shipments do not outperform CLP ES for the given waiting time restrictions, but this happens if we further increase $W_1{}^{max}$. Under the mentioned conditions, the base stock levels with CLP ES tend to be high to avoid expensive emergency shipments.

Obviously, the combined policy works best. Still the additional gain is surprisingly large: it exceeds the combined savings of the individual policies. The reason is that under CLP ES, the mean waiting time *for class 2 customers* tends to be considerably below the target; the class 1 target is usually the bottleneck. By including selective emergency shipments, we are able to push the actual performance of low priority customers closer to the target (leading to deviations of 0.04% instead of 29% in the experiments with 100 items).

7. Conclusions and discussion

We discuss our main findings and the complexity of various model extensions.

7.1. Conclusions

First, we developed two heuristics that are accurate (average gaps to the lower bound well below 1%) and fast. Clearly, greedy approaches are not necessary to find good solutions: integer programming with limited columns is simple and works well. Second, selective emergency shipments have clear added value, with average savings of 4.4% compared to one-size-fits-all policies. The approach also outperforms critical level policies when holding costs are low, premium waiting times are not very tight, and regular shipment times are short. Then, emergency shipments are very expensive and should thus be avoided. Differentiation through selective emergency shipments is most useful for expensive slow movers, since the approach has most impact when little or no stock is kept of an item. Finally, we find large savings (13.9% on average) by jointly using critical levels and selective emergency shipments for differentiation in spare parts supply.

7.2. Model extensions

One obvious extension option is **to consider more than two customer classes:** as the analysis of an item policy is separate from the optimization of these policies, we can easily extend the model to more customer classes, provided that we have at most two classes per item policy: we require a new column generation procedure, but the evaluation of an item policy does not change. Kranenburg and Van Houtum (2008) find relatively large savings by smartly assigning five customer groups over 2 main classes, with the assignment being the same for all item policies. By allowing this assignment to differ among item policies, we expect that even larger savings are possible. If we use backordering for at most 1 customer class, we can also analyze item policies with more than 2 classes: the analysis then still remains simple, even if critical levels are positive. Otherwise, each additional class adds an additional dimension to the Markov chain, which complicates the analysis. If critical levels are zero, we can find waiting times by aggregating classes. However, if critical levels are positive and differ per class, aggregation is not possible.

We also see simple options for extensions to **customers with multiple systems that might have common components:** instead of one waiting time restriction per customer class, we now need a restriction for each combination of customer class and system type. Separate restrictions per system type are necessary if an item found in multiple systems has different multiplicities per system. Then, that item's waiting time will influence each system's waiting time differently. These new restrictions will result in slight changes to the reduced cost function of Section 4.2. Still, the solution approach will not change greatly: the methods for analyzing item policies and finding policies with negative reduced costs will not change.

Besides these simple extensions, we see more promising research areas. One area is the use of better shipment strategies: by also considering the system state when selecting a shipment option, further savings might be possible (see Section 3.2). Another area is that of multi-echelon systems: in such models, we explicitly need to stock items at some location to meet emergency requests, resulting in additional costs. Finally, we see selective lateral transshipments between warehouses as an additional promising differentiation option.

Table A1
Performance comparison under exponential and deterministic lead times.

Shipment strategy	Average deviation		Maximum deviation	
	EW_1(%)	EW_2(%)	EW_1(%)	EW_2(%)
Em. shipments for both classes	0.3	0.1	1.7	0.6
Em. shipments for premium customers only	1.7	0.3	9.2	0.9
Backordering for both classes	8.8	0.7	14.1	1.5

Appendix A. Sensitivity to lead time distribution

For a single-class system with emergency shipments, Alfredsson and Verrijdt (1999) show that the performance measures do not depend on the distribution of the lead time. We therefore test whether this observation still holds for a two-class system. For 48 problem instances, we used simulation to find mean waiting times per class for both deterministic and exponential lead times. The regular shipment time is 5 or 10 days, the emergency shipment time is 1 or 2 days, and the class demand rates (m_1;m_2) are either (0.01;0.04) or (0.1;0.4).

Per shipment strategy, we compute the waiting time deviations as a fraction of the exponentially distributed times. Table A1 shows the results for cases with waiting times of at least 10^{-3}. We find that the lead time distribution still has little influence on the waiting times: in general, the average deviations are small. The deviations increase as backordering is used for more classes, particularly for class 1 waiting times, but the differences remain reasonable even under full backordering (we find deviations of 14% when EW_1 is 1.5 and 1.7, respectively). Also, in practice we expect waiting times under backordering to show more variability than those under emergency shipments, especially when items are repaired (as is common for expensive slow movers). Under full backordering, we thus expect exponential shipment times to characterize the supply process more accurately than deterministic times.

References

Alfredsson, P., Verrijdt, J., 1999. Modeling emergency supply flexibility in a two-echelon inventory system. Management Science 45, 1416–1431.
Alvarez, E.M., Van der Heijden, M.C., and Zijm, W.H.M. (2010). The selective use of emergency shipments for service-contract differentiation. BETA Working paper 322, BETA research school, The Netherlands, http://beta.ieis.tue.nl/publications.
Arslan, H., Graves, S.C., Roemer, T.A., 2007. A single-product inventory model for multiple demand classes. Management Science 53, 1486–1500.
Cohen, M.A., Agrawal, N., Agrawal, V., 2006. Winning in the aftermarket. Harvard Business Review 84, 129–138.

De Véricourt, F., Karaesmen, F., Dallery, Y., 2002. Optimal stock allocation for a capacitated supply system. Management Science 48, 1486–1501.

Dekker, R., Hill, R.M., Kleijn, M.J., Teunter, R.H., 2002. On the $(S-1, S)$ lost sales inventory model with priority demand classes. Naval Research Logistics 49, 593–610.

Gayon, J.-P., De Véricourt, F., Karaesmen, F., 2009. Stock rationing in an M/Er/1 multi-class make-to-stock queue with backorders. IIE Transactions 41, 1096–1109.

Ha, A.Y., 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. Management Science 43, 1093–1103.

Ha, A.Y., 1997b. Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. Naval Research Logistics 44, 457–472.

Ha, A.Y., 2000. Stock rationing in an M/Ek/1 make-to-stock queue. Management Science 46, 77–87.

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

Kranenburg, A.A., Van Houtum, G.J., 2007. Effect of commonality on spare parts provisioning costs for capital goods. International Journal of Production Economics 108, 221–227.

Kranenburg, A.A., Van Houtum, G.J., 2008. Service differentiation in spare parts inventory management. Journal of the Operational Research Society 59, 946–955.

Möllering, K.T., Thonemann, U.W., 2010. An optimal constant level rationing policy under service level constraints. OR Spectrum 32, 319–341.

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.

Topkis, D.M., 1968. Optimal ordering and rationing policies in a nonstationary dynamic inventory model with n demand classes. Management Science 15, 160–176.

Van Utterbeeck, F., Wong, H., Van Oudheusden, D., Cattrysse, D., 2009. The effects of resupply flexibility on the design of service parts supply systems. Transportation Research Part E: Logistics and Transportation Review 45, 72–85.

Veinott, A.F., 1965. Optimal policy in a dynamic, single product, nonstationary inventory model with several demand classes. Operations Research 13, 761–778.

Wong, H., Van Oudheusden, D., Cattrysse, D., 2007. Two-echelon multi- item spare parts systems with emergency supply flexibility and waiting time constraints. IIE Transactions 39, 1045–1057.