OR Spectrum (2018) 40:159–186
https://doi.org/10.1007/s00291-017-0497-8

REGULAR ARTICLE

Transshipment policies for systems with multiple retailers and two demand classes

Zümbül Atan[1] · **Lawrence V. Snyder**[2] · **George R. Wilson**[2]

Received: 22 September 2016 / Accepted: 16 November 2017 / Published online: 24 November 2017

Abstract Many retailers discriminate among their customers based on their value to the firm. Instead of losing a customer due to this discrimination or lack of inventory, a retailer might prefer to place a transshipment request with other retailers to satisfy the customer's demand. The system as a whole can benefit from this type of transshipments. In this paper, we study a problem of a centrally controlled system with multiple retailers. Each retailer serves two types of customers: high priority and low priority. Each retailer employs a rationing policy with a rationing level k in the context of a continuous-review (r, Q) inventory replenishment model. The overall policy is referred to as an (r, k, Q) policy. Retailers can transship items from either other retailers or a more expensive central depot. We propose an enumeration-based approximation to find the cost-minimizing policy parameters for the individual retailers and an approximation procedure to solve the combined rationing and transshipment problem. The latter relies on adjusting the demand arrival rates and the unit transshipment costs for both types of customers at all retailers. An extensive numerical study highlights the impact of transshipments on the retailers' rationing policies. Without transshipment opportunities among each other, retailers set their policy parameters so

✉ Zümbül Atan
Z.Atan@tue.nl

Lawrence V. Snyder
larry.snyder@lehigh.edu

George R. Wilson
grw3@lehigh.edu

1 Department of Industrial Engineering and Innovation Sciences, School of Industrial Engineering, Eindhoven University of Technology, P.O.Box 513, Pav.E8, 5600 MB Eindhoven, The Netherlands

2 Department of Industrial and Systems Engineering, Lehigh University, 200 West Packer Ave., Mohler Lab, Bethlehem, PA 18015, USA

that the resulting service levels are high for both types of customers. Allowing transshipments results in more aggressive rationing policies, and retailers with aggressive rationing policies benefit from the transshipments the most.

Keywords Rationing · Transshipment · Approximation procedure

1 Introduction

Practitioners and academics have proposed multiple strategies for achieving significant cost and service benefits that increase the competitiveness of multi-echelon inventory systems. Some of these benefits can be achieved by physical pooling of inventories (Eppen 1979). As demands across different locations are aggregated, it becomes more likely that a high demand from one customer is offset by a low demand from another. A similar and commonly employed strategy is to allow the movement of inventories among the locations belonging to the same echelon. These stock movements are called lateral transshipments.[1] Transshipments are similar in the sense that they allow stock sharing, but without the need to physically stock inventories in the same location. While risk pooling represents physical integration, transshipments are often considered virtual integration through information sharing. Transshipments provide an efficient mechanism for correcting mismatches between locations demands and their available inventories. They can improve customer service levels and lead to cost reductions without the need to increase stock.

Most models in the supply chain literature assume that the demands arising from different sources are handled in the same way. In practice, however, for a single product, different customers may have different stockout costs or service-level requirements, or may be of different importance to the supplier. Therefore, it can be appropriate to distinguish between classes of customers and use different policies in satisfying their demands. Transshipments can be especially helpful in this case. When the retailer chooses not to satisfy the demands of its low-priority customers due to low inventory levels, it might request another retailer in its supply chain to satisfy these customers' demands via transshipment. Although transshipping can be expensive, it might be more expensive not to satisfy a customer's demand, even if it is of low priority.

Multiple demand classes occur naturally in many inventory systems. Consider, for example, a spare parts system. In such a system a part may be required for the production of multiple products, but may be more crucial for one product than for others. Here, products can be viewed as different demand classes. Another example is a firm that has some customers who have a contract for its products and are guaranteed a certain service level, as well as "off-the-street" customers to whom the firm has no such obligation.

The easiest policy to use for a single-location system with multiple demand classes is simply to hold a separate inventory for each demand class. Although this policy is easy to implement in practice, it does not take advantage of the risk-pooling effect (Eppen 1979) and would therefore result in no inventory savings. On the other hand,

[1] In the rest of the paper, we refer to "lateral transshipments" as "transshipments."

using the same pool of inventory to satisfy demands from different customer classes without differentiating them would result in increased inventory costs, since the highest required service level is generally used to determine inventory levels. *Rationing* lies between these two cases.

Rationing, which we also refer to as a *critical-level policy*, is defined as the practice of reserving part of the stock for high-priority demand classes (Melchiors et al. 2000). It has been demonstrated that rationing is an effective way of handling different demand classes with different stockout costs and service-level requirements (Kleijn and Dekker 1998). In a rationing policy, when the inventory level drops to a certain point, called the *critical level*, demands from lower-priority customers are no longer satisfied. This policy can be extended to more than two customer demand classes by considering multiple critical levels.

In general, an optimal critical-level policy should take into account the remaining time until the arrival of the next replenishment order. The reason is that, if the inventory level is below the critical level, but it is known that a replenishment order will arrive within a short period of time, it may not be optimal to refuse a low-priority demand arrival, especially if the probability of a high-priority demand arrival within the remaining lead time is very small. However, from a practical point of view, employing such a dynamic rationing policy would be extremely difficult. We study a static rationing policy without taking the remaining time of the arrival of the next replenishment into account.

We consider a multi-retailer system in which each retailer has two customer demand classes. The demand of each customer class at each retailer has a Poisson distribution. Each retailer[2] employs a critical-level policy in the context of a continuous-review (r, Q) inventory model. This policy is referred to as an (r, k, Q) policy (Nahmias and Demmy 1981), where r is the reorder point, k is the critical level, and Q is the order quantity. Whenever a retailer is out of stock or cannot satisfy a demand because of its rationing policy, it can request an item from another retailer in the system and satisfy the demand of its customer. If this retailer is out of stock as well, the original retailer places a transshipment request with another retailer, and so on. We assume that each retailer has a fixed sequence of preferred retailers from which to request transshipments. This sequence can be based, for example, on the distance between the retailers. We assume a static sequence, i.e., it does not depend on retailers' inventory levels or other state variables. If none of the retailers can satisfy the transshipment request, it is satisfied by a central depot. We refer to this type of shipment from the depot also as a transshipment, distinguishing it from a replenishment shipment. The transshipping location sends the item directly to the customer, and the demand is assumed to be satisfied as soon as the item is shipped. The system incurs a cost for each unit transshipped. The aim of a centralized decision maker is to determine the policy parameters for all retailers with an objective of minimizing the total cost of the system.

The main motivation behind this study is the increased need for companies facing fierce competition to satisfy all of their customers' demands. It is important to satisfy not only the demand of high-priority customers, but also the demand of low-priority

[2] We use the term "retailer," but the model could equally well apply to other echelons.

customers who may constitute a higher percentage of the total demand. This research can benefit many companies in the automotive, high-tech and apparel industries. It can especially benefit *online-retailing companies* with multiple stock-keeping locations since customers typically have no preference regarding the location from which their goods are shipped, as long as they arrive when promised. For online-retailing companies, high-priority customers are the ones who pay more and request fast shipments, while low-priority customers pay less and accept to wait longer for the shipment of their products. In this context, independent of the location satisfying her demand, a high-priority customer always has high priority and a low-priority customer always has low priority. The benefits of transshipments are heightened by recent regulations that force Amazon and similar online-retailing companies to collect sales tax in more and more states in the USA. This makes them less price competitive. According to articles in DailyFinance (Brownell 2013) and Slate (Manjoo 2012), Amazon's response has been to construct distribution centers in these states and promise 1-day delivery to match what local brick-and-mortar merchants can do. For example, a high-end camera may be offered locally by some merchant and Amazon must work hard to keep its one-day delivery pledge to be competitive. Rationing comes in when the item is one for which Amazon finds itself in a highly competitive situation in some state. That particular Amazon distribution center will not transship this item elsewhere after reaching a critical level needed to stay competitive locally.

Motivated by these examples, we *provide an approximation procedure* to solve transshipment problems, which involve multiple stock-keeping locations receiving demands from two customer types. Through our numerical analysis, we provide *insights* into how transshipments affect the inventory replenishment policies of retailers. Our insights help company managers in making decisions on how to make use of transshipments in reducing their operational costs. More specifically, depending on the retailers' pre-transshipment inventory replenishment policies, managers can design a transshipment policy for guaranteeing substantial cost reductions. The design should tell whether high- and low-priority customer demands have the same or different priorities at the other retailers. It should also tell whether it is enough to transship high-priority customer demands only. This study provides insights to help in such design decisions.

The remainder of this paper is organized as follows: In Sect. 2, we provide a brief review of the literature on transshipment and rationing policies. In Sect. 3, we introduce the notation, define the problem and summarize our solution procedure. We explain the single-retailer problem and our enumeration-based approximation to solve it in Sect. 4. We introduce our approximation procedure for the multi-retailer problem in Sect. 5. In Sect. 6, we summarize the results of our numerical analysis. Finally, we present our conclusions and suggest future research directions in Sect. 7.

2 Literature review

The research presented in this paper considers two key issues: (i) transshipments and (ii) inventory rationing. In this section, we briefly review the existing literature in these areas.

2.1 Transshipments

Typically, lateral transshipment models are divided into two categories: reactive (emergency) transshipments and proactive (preventive) transshipments. Reactive transshipments refer to transshipment requests made by a retailer with no stock from another retailer with positive stock. This type of transshipment responds to stockouts. On the other hand, proactive transshipments imply redistribution of the stock among retailers as a result of anticipated future stockouts; retailers do not need to wait until one has no stock to transship. In this paper, we study a continuous-review policy with reactive transshipments. The transshipping location does not hold back any inventory. Next, we briefly summarize the most relevant literature on reactive transshipments. The interested readers are referred to Paterson et al. (2011) for an extensive literature review on periodic-review inventory models and continuous-review inventory models with proactive transshipments.

An important problem in the transshipment literature is the determination of the retailer to transship from. In this paper, we consider the unit transshipment cost to decide which retailer to transship from. The same selection criterion is studied in multiple papers (Kukreja et al. 2001; Archibald 2007). In an early work Lee (1987) tests several other rules for selecting the retailer to transship from, including using random selection, maximum stock on hand and smallest number of outstanding orders. He reports no significant differences in the performances of the three selection rules and concludes that, independent of the selection rule, reactive transshipments reduce total system costs. The same is concluded in Seidscher and Minner (2013). Another option is the distance-based transshipment rule, which implies that the items are transshipped from a nearby location with adequate supply (Hu et al. 2005; Kukreja and Schmidt 2005). Axsäter (1990) assumes transshipments from a randomly selected location that has stock on hand at the moment the request is placed. The random selection rule is commonly assumed in the literature. In fact, Comez-Dolgan and Fescioglu-Unver (2015) state that this rule performs quite well. Archibald et al. (2009) suggest a different selection criterion based on the fair charge for the transshipped inventory. They report substantial cost savings compared to other policies.

The standard (r, Q) policy is widely used in the transshipment literature. Needham and Evers (1998) provide a mathematical tool as an aid to make transship or do-not-transship decisions. Similarly, Evers (2001), Axsäter (2003b), Axsäter (2003a), Minner et al. (2003), Xu et al. (2003), Minner and Silver (2005), Ching et al. (2003) and Olsson (2009) study (r, Q) policies with an objective of providing rules for transshipment decisions.

Most of the studies on reactive transshipments assume that transshipments only happen when a stockout occurs. In our paper, the transshipments for low-priority demands occur even if all locations have positive stocks. However, even in a setting with a single customer type, it might be optimal to request transshipments whenever the inventory hits a threshold level (Zhao et al. 2006; Grahovac and Chakravarty 2001). The resulting models are quite difficult to analyze due to the additional decision variable. In fact, Archibald et al. (1997) prove that the optimal transshipment policy has two threshold-level characteristics. They show that (i) if it is optimal to satisfy a transshipment request at a given stock level, it is also optimal to do the same at higher

stock levels and (ii) if at a given time it is optimal to satisfy a transshipment request, it is also optimal to do the same at times that are closer to the next replenishment moment.

In this paper we assume that each transshipment request is for a single item. Given that transshipments are not for free, it might be cost-effective to transship multiple units in anticipation of future shortages. These are called hybrid transshipments since they are both reactive and proactive. There are a limited number of papers that study hybrid transshipments (Paterson et al. 2012; Glazebrook et al. 2015).

As discussed in Sect. 1, our study can especially benefit online-retailing companies with multiple stock-keeping locations. Yang et al. (2014) also consider an online-retailing company with dynamic demand and develop heuristic transshipment strategies. In a related work, Torabi et al. (2015) study an online-retailing company and, for specific demand data, develop a model to determine the optimal order-delivery plan that minimizes the transportation and transshipment costs.

Another application field for our problem/model is the spare part inventory systems that serve installed bases of advanced machines. In the transshipment literature, there exist multiple studies that are motivated by spare part inventory systems. Examples include Kranenburg (2006), Tiacci and Saetta (2011), van Wijk et al. (2012), Yang et al. (2013) and Olsson (2015). Our model differs from these examples and therefore contributes to this field by differentiating among spare part demands, i.e., some demands for spare parts have higher priority than others.

This paper is also related to previous work on dual-channel retail systems where retail stores (direct channel) serve in-store customers and use excess stock to fill some online orders (indirect channel). Serving online orders through excess stock implies transshipments, but in contrast to our work, transshipments happen in one direction; hence, they are asymmetric. Researchers suggest significant profit gains, improved service levels and reduced inventories through integration and coordination of direct and indirect channels (Seifert et al. 2006; Liang et al. 2014; Zhao et al. 2016).

2.2 Inventory rationing

Problems involving several demand classes and the concept of a "critical-level policy" was first introduced by Veinott (1965). Subsequently, Topkis (1968) proves the optimality of this policy for both backordering and lost sales cases. Topkis (1968) argues that optimal critical levels are decreasing functions of the time remaining until the next replenishment request. Kaplan (1969) and Evans (1968) obtain the same results as Topkis (1968), independently, for two customer demand classes. Following these seminal works, many authors study the critical-level policy under different settings. We refer to Teunter and Haneveld (2008) for a thorough review. In this section we give a brief review on the latest articles which are most relevant to our study.

In this paper, we study an (r, Q) inventory policy with a static critical level k. This policy has been studied extensively in the literature. Assuming at most one outstanding order, Nahmias and Demmy (1981) derive expressions for the expected backorders and service levels for each demand class. The same problem with multiple outstanding orders is studied by Deshpande et al. (2003). The authors design a threshold clearing

mechanism and develop an efficient solution algorithm for computing policy parameters. The same clearing mechanism is studied by Wang et al. (2013). Fadiloglu and Bulut (2010) handle the problem with multiple outstanding orders by adjusting the inventory level by including outstanding orders.

In our study, each retailer employs an (r, k, Q) policy with lost sales. The same policy is studied by Melchiors et al. (2000), Wang et al. (2015) and Isotupa and Samanta (2013). Assuming constant replenishment lead times, Melchiors et al. (2000) propose an exact optimization procedure to find the optimal reorder and critical levels. Wang et al. (2015) prove that the optimal rationing policy is a combination of a static policy before order release and a dynamic policy during the replenishment lead time. Isotupa and Samanta (2013) extend the results in Melchiors et al. (2000) by assuming arbitrarily distributed lead time.

Fadiloglu and Bulut (2010) show that a dynamic rationing policy, which allows the critical level to change, can outperform a constant critical-level policy. The interested reader can refer to Hung et al. (2012), Chew et al. (2013), Wang and Tang (2014) and Liu et al. (2015) for recent findings on dynamic rationing policies.

2.3 Contribution

Our research contributes to the literature on rationing and transshipments by providing a formulation for multiple-retailer, multiple-demand-class problems with rationing and transshipment policies and introducing an approximation procedure for solving these problems. To the best of our knowledge, the combined problem of rationing and transshipments has been studied previously only by Alvarez et al. (2014). Alvarez et al. (2014) assume lateral transshipments for high-priority customers only. However, companies facing fierce competition have an increased need to satisfy not only the demand of high-priority customers, but also the demand of low-priority customers who may constitute a higher percentage of the total demand. Based on this need, we allow transshipments for both types of customers and provide an approximation procedure to solve transshipment problems. In addition, through our numerical analysis, we provide insights into how transshipments affect the inventory replenishment policies of retailers who ration their customers' demands.

Our model allows transshipments among any number of retailers. In fact, when a retailer cannot satisfy its customer's demand either due to stockouts or because its inventory level is below the critical level, it places a transshipment request either with other retailers in the system or with the central depot. Hence, our model enables practitioners to employ rationing policies without losing their customers. Our approximation procedure relies on solving single-retailer problems without considering transshipments and then linking these problems by redirecting the unsatisfied demands to other locations.

3 Model framework

In this section, we introduce the notation and define our problem.

3.1 Notation and preliminaries

We consider a centrally controlled, single-item continuous-review inventory system with one central depot and N retailers. We use index i to represent the retailer. If not stated otherwise, the following definitions apply for all $i \in \{1, 2, \ldots, N\}$. Each retailer has two types of customers: high priority and low priority. Index j represents the customer type, with $j = 1$ being high priority and $j = 2$ being low priority. Retailer i follows a continuous-review (r_i, k_i, Q_i) policy, which operates as follows: Whenever the inventory level drops to the reorder level r_i, a replenishment order of size Q_i is placed with the central depot. The replenishment lead time T_i is constant. Both classes' demand is satisfied whenever the inventory level exceeds or equals the critical level, k_i. If the inventory level is less than k_i, only high-priority demand is satisfied from the stock on hand and low-priority demand is satisfied through transshipments. If the inventory level is zero, both classes' demand is satisfied through transshipments.

Let $D_{i,j}(t)$, $j = 1, 2$, be a random variable denoting the demand from type-j customers during t time units. $D_i(t) := D_{i,1}(t) + D_{i,2}(t)$ is the total demand at retailer i during t time units. We assume unit Poisson demand with arrival rate $\lambda_{i,j}$, $j = 1, 2$. The total demand rate is $\lambda_i = \lambda_{i,1} + \lambda_{i,2}$. Defining $\mathbb{E}[X]$ as the expectation of the random variable X, we have $\mathbb{E}[D_{i,j}(t)] = \lambda_{i,j}t$ and $\mathbb{E}[D_i(t)] = \lambda_i t$.

Retailer i's cost of not satisfying a demand from demand class j directly from its own stock is $\pi_{i,j}$, with $\pi_{i,1} \geq \pi_{i,2} > 0$. For retailer i, the fixed ordering cost is K_i and the unit holding cost per unit time is $h_i > 0$.

Similar to Hadley and Whitin (1963), Nahmias and Demmy (1981) and Melchiors et al. (2000), we restrict ourselves to policies in which there is at most one outstanding replenishment order. This implies that at the time a replenishment order is placed, the net inventory and the inventory position are identical. A sufficient condition to ensure that at most one order is outstanding is $r_i < Q_i$. This condition ensures that when an outstanding order arrives, the inventory level goes above r_i, which implies that another outstanding order does not exist. In addition, we require $k_i < Q_i$, for the model to be tractable. We note that, for the problem under consideration, although we use an (r, k, Q) policy, the optimal policy is unknown. It is possible to combine other inventory replenishment policies with the critical-level policy. However, given that the (r, Q) policy is preferred and widely used in practice and is, in fact, the optimal inventory replenishment policy under our assumptions (Axsäter 2015), we assume the (r, k, Q) inventory and rationing policy.

Next, we clarify the transshipment process. When retailer $i \in \{1, 2, \ldots, N\}$ cannot satisfy a demand immediately, it can place a transshipment request with other retailers. Let O_i be the ordered list of retailers that retailer i can request transshipments from. Hence, O_i is a vector with $N - 1$ elements indicating the rankings. We have $O_i = [o_i(n)]_{n=1}^{N-1}$, where $o_i(n)$ is retailer i's nth choice of retailer to transship from. For example, if $N = 3$ and $O_2 = [o_2(1), o_2(2)] = [3, 1]$, in case of a stockout, retailer 2 initially considers requesting an item from retailer 3. If the demand is from a high-priority customer, retailer 3 satisfies the demand as long as he has positive inventory. If, on the other hand, the demand is from a low-priority customer, retailer 3 satisfies the demand if his inventory level is above k_3. If retailer 3 cannot satisfy retailer 2's demand, retailer 2 places a transshipment request with retailer 1. If all retailers fail

to satisfy the demand, a transshipment request is placed with the central depot. We assume that the central depot always has enough inventory to satisfy all transshipment requests. Transshipments from both retailers and the central depot are instantaneous. The unit cost of satisfying retailer i's type-j demand via transshipment from retailer $o_i(n)$, referred to as the "transshipment cost," is $c_{i,j,o_i(n)}$, and the unit cost of satisfying it from the central depot is $c_{i,j,d}$. Transshipments among retailers are cheaper than transshipping from the central depot, i.e., $c_{i,j,d} > c_{i,j,o_i(n)}, \forall n \in \{1, 2, \ldots, N-1\}$. When companies group the retailers among which transshipments are allowed they rely on distance. The central depot is typically far from most groups. Therefore, it costs more to ship an item from the depot to a customer who normally needs to be served from one of the retailers, which is located closer to the customer. We assume that the sequence O_i is determined based on the unit transshipment costs. Our approximation procedure is flexible enough to handle any other static sequencing rule.

If transshipments among retailers are not allowed, each retailer can only place transshipment requests with the central depot. The unit cost of transshipping retailer i's type-j demand from the central depot equals retailer i's cost of not satisfying a type-j demand. Hence, without any transshipment opportunities among the retailers, we have $c_{i,j,d} = \pi_{i,j}, \forall i \in \{1, 2, \ldots, N\}$ and $j = 1, 2$. For the case where transshipments among the retailers are allowed we propose an approximation procedure to calculate the unit cost of not satisfying the demand directly from stock. (Refer to Sect. 5.)

We assume that a type-j demand at retailer i is again a type-j demand for the retailers in the rest of the system. In reality, a high-priority customer of one retailer might be a low-priority customer for another retailer. Our approximation procedure can be easily modified to handle this case, but we make the simpler assumption for ease of exposition. (Refer to Sect. 6.5.1 for the modified case.)

3.2 Problem definition

We let $\mathbf{r} := [r_i]_{i=1}^N$, $\mathbf{k} := [k_i]_{i=1}^N$ and $\mathbf{Q} := [Q_i]_{i=1}^N$ be the vectors of retailers' reorder levels, critical levels and order quantities, respectively, and define the following long-run averages:

- $I_i(\mathbf{r}, \mathbf{k}, \mathbf{Q})$ the average inventory level at retailer i,
- $A_i(\mathbf{r}, \mathbf{k}, \mathbf{Q})$ the average number of replenishment orders placed by retailer i per unit time,
- $S_{i,j,o_i(n)}(\mathbf{r}, \mathbf{k}, \mathbf{Q})$ the average number of type-j demands of retailer i satisfied via transshipment from retailer $o_i(n)$ per unit time,
- $S_{i,j,d}(\mathbf{r}, \mathbf{k}, \mathbf{Q})$ the average number of type-j demands of retailer i satisfied via transshipment from the central depot per unit time.

The problem is to determine the optimal stocking and rationing policies of all the retailers to minimize the total expected cost. Defining $C(\mathbf{r}, \mathbf{k}, \mathbf{Q})$ as the total expected cost, we express our problem as:

$$\min C(\mathbf{r}, \mathbf{k}, \mathbf{Q}) = \sum_{i=1}^{N} \left[h_i I_i(\mathbf{r}, \mathbf{k}, \mathbf{Q}) + K_i A_i(\mathbf{r}, \mathbf{k}, \mathbf{Q}) + \sum_{j=1}^{2} \left(c_{i,j,d} S_{i,j,d}(\mathbf{r}, \mathbf{k}, \mathbf{Q}) \right. \right.$$

$$+\sum_{n=1}^{N-1} c_{i,j,o_i(n)} S_{i,j,o_i(n)}(\mathbf{r},\mathbf{k},\mathbf{Q})\Big)\Big]$$

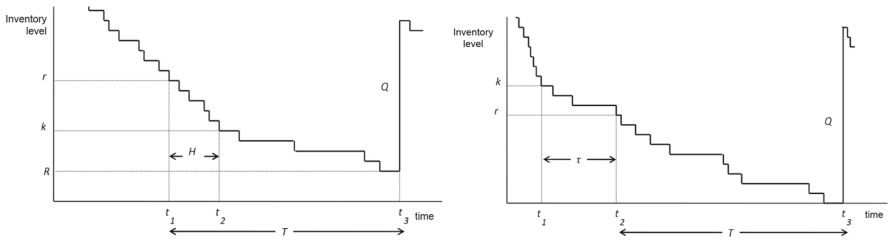
$$\text{s.t.} \quad r_i < Q_i \quad \forall i \in \{1,2,\ldots,N\}$$
$$k_i < Q_i \quad \forall i \in \{1,2,\ldots,N\}$$
$$r_i \geq 0,\ k_i \geq 0,\ Q_i \geq 0 \quad \forall i \in \{1,2,\ldots,N\}. \tag{1}$$

(1) is a nonlinear optimization problem. The long-run averages defined in Sect. 3.2 depend on policy parameters **r**, **k** and **Q**. These dependencies make the problem difficult to solve to optimality. This is why we propose *an approximation procedure*, which relies on solving single-retailer problems without considering transshipments and then linking these problems by redirecting the unsatisfied demands to retailers in $O_i, \forall i \in \{1,2,\ldots,N\}$. The details of the single-retailer problem (SiReP) and the approximation procedure for the multi-retailer problem (MuReP) are in Sects. 4 and 5, respectively.

4 Single-retailer problem (SiReP)

Melchiors et al. (2000) consider a single-retailer problem with two demand classes. Their assumptions are the same as our assumptions for individual retailers. They study the (r, k, Q) policy and provide a procedure to optimize r and k. In this section, we provide a brief explanation of the procedure and outline our contribution to it.

For ease of exposition, we drop the index i. We define $X(t)$ as the on-hand inventory level at time t. The corresponding stochastic process $\{X(t), t \geq 0\}$ is a regenerative process. By defining a cycle as the time between two consecutive replenishment order requests, we can use the renewal-reward theorem to find the average cost per unit time. We consider two different scenarios with $k < r$ and $k \geq r$. Figure 1a, b depicts the behavior of $X(t)$ when $k < r$ and $k \geq r$, respectively. When $k < r$, the retailer starts to reject the low-priority customers' demands after placing a replenishment order. In Fig. 1a, H is the time from placing a replenishment order until the time when the low-priority customers' demands start to be rejected. H has an Erlang distribution with parameters $r - k$ and λ. When $k \geq r$, the retailer starts to reject the low-priority customers' demand before placing a replenishment order. In Fig. 1b,



(a) $k < r$ **(b)** $k \geq r$

Fig. 1 Inventory process

Table 1 Expressions for L_1, L_2 and CL

	$k < r$	$k \geq r$
L_1	$\mathbb{E}_{D_1(T-H)}\big[D_1(T-H)-k\big]^+$	$\mathbb{E}_{D_1(T)}\big[D_1(T)-k\big]^+$
L_2	$\mathbb{E}_{D_2(T-H)}\big[D_2(T-H)\big]$	$\mathbb{E}_{D_2(T+\tau)}\big[D_2(T+\tau)\big]$
CL	$T+\frac{Q+\mathbb{E}[R]-r}{\lambda}$	$T+\frac{Q+\mathbb{E}[R]-k}{\lambda}+\frac{k-r}{\lambda_1}$

$\tau := \{t \geq 0 : D_1 \geq k-r\}$ is the time between the start of rejecting low-priority customers and placing an order. We have $\mathbb{E}[\tau] = \frac{k-r}{\lambda_1}$. R stands for the inventory level just before a replenishment order arrives. For both cases R takes values between 0 and r. The distribution of R is

$$\mathbb{P}(R=w) = \begin{cases} \mathbb{P}(D_1(T-H) \geq k), & \text{for } w=0, \\ \mathbb{P}(D_1(T-H)=k-w), & \text{for } 0 < w \leq k, \\ \mathbb{P}(D(T)=r-w), & \text{for } k < w \leq r. \end{cases}$$

This distribution can be simplified when $k \geq r$ since for this case the hitting time H is not defined. The point of equality where $k = r$ is not a special point. We include the equality in $k > r$-type policy and write $k \geq r$ since H is not defined for $k = r$ or $k > r$. Using the distributions and necessary moments of the random variables H, τ and R, it is straightforward to write exact expressions for the cycle costs, cycle lengths and, therefore, the average costs.

A particular performance measure that plays an important role in the multi-retailer problem is the expected proportion of demands met directly from stock. Define β_j as the expected proportion of type-j demands met directly from stock. The percentage of demands from class j that cannot be met immediately from stock is given by $\frac{L_j}{\lambda_j \mathrm{CL}}$, where L_j is the expected number of unmet demands per cycle for type-j customers and CL is the expected cycle length. Hence, $\beta_j = 1 - \frac{L_j}{\lambda_j \mathrm{CL}}$. In Table 1 we provide the formulas for L_1, L_2 and CL.

Our numerical analysis shows that the evaluation of a single (r, k, Q) policy using the method by Melchiors et al. (2000) is quite time-consuming because it involves numerical integration. Our solution for the overall problem with multiple retailers relies on solving the single-retailer problem multiple times. In order to facilitate the procedure, we replace the random variable H with its expectation, $\mathbb{E}[H] = \frac{r-k}{\lambda}$, thereby removing the need for integration over the random variable H in the calculation of the policy parameters. This approximation only affects the analysis for the case with $k < r$. The average cost is affected slightly by this approximation. Based on our numerical analysis, we observe that optimal policy parameters, the calculations of which are explained next, are not affected much. (See Sect. 6.1.)

The optimization procedure by Melchiors et al. (2000) relies on enumeration and bounding. The authors assume that Q is given, and they provide an algorithm to find the corresponding optimal values of r and k, denoted $r^*(Q)$ and $k^*(Q)$. Given that $D(T)$ is the total demand during the replenishment lead time, T, they prove an upper bound for $r^*(Q)$ given by

$$\bar{r}(Q) = \min\left\{ r \geq 0 : \mathbb{P}(D(T) \geq r+1) \leq \frac{h}{h + \frac{(\pi_1\lambda_1 + \pi_2\lambda_2)}{Q}} \right\}.$$

In order to find $r^*(Q)$, we need to enumerate all possible values from 0 to $\bar{r}(Q)$. In addition, an enumeration is needed to find the cost-minimizing value of k, satisfying $k(Q) < r^*(Q)$. Let this value be $k^1(Q)$. When $k \geq r^*(Q)$, for fixed values of r and Q, the average cost function is either convex or concave in k and it is easy to find k to minimize $\mathrm{TC}^{k \geq r}$, i.e., the total average cost when $k \geq r$. Let $k^2(Q)$ be this value. $k^*(Q)$ is $k^1(Q)$ if $\mathrm{TC}^{k^1(Q) < r^*(Q)} < \mathrm{TC}^{k^2(Q) \geq r^*(Q)}$, and $k^*(Q)$ is $k^2(Q)$, otherwise.

In contrast to Melchiors et al. (2000), we replace H by $\mathbb{E}[H]$ and use enumeration over Q to find the values of r, k and Q to minimize the approximated cost. We refer to this approximate solution for the single-retailer problem as *an enumeration-based approximation*. Define $Q^*(r,k)$ as the optimal value of Q for given values of r and k. We use 1 and $Q^*(0,0)$ as the lower and upper bounds for Q, respectively. For each value of Q, we find $r^*(Q)$ and $k^*(Q)$ and calculate the average cost, $C(r^*(Q), k^*(Q), Q)$. We choose the solution with the lowest average cost as our final solution.

5 Multi-retailer problem (MuReP)

Given that we know how to solve the problems of individual retailers, we now introduce our *approximation procedure* to solve the overall problem with transshipment opportunities among retailers.

The first step of the procedure is to solve N single-retailer problems without transshipments by using the enumeration-based approximation for SiReP described in Sect. 4 to find the policy parameters r_i, k_i and $Q_i, \forall i \in \{1, 2, \ldots, N\}$.

In the second step of the procedure, we introduce the transshipments and we adjust the demand rates and the penalty costs. Define $\gamma_{i,j,o_i(n)}$ as the expected demand rate of type-j items retailer i requests from retailer $o_i(n)$. As defined in Sect. 3.2, $o_i(n)$ is the retailer that is in the nth position in the ordered list O_i. $\gamma_{i,j,o_i(n)}$ is calculated as follows:

$$\gamma_{i,j,o_i(n)} = (1-\beta_{i,j})\lambda_{i,j}\left[\Pi_{m=1}^{n-1}(1-\beta_{o_i(m),j})\right]\beta_{o_i(n),j} \quad \forall i \in \{1,2,\ldots,N\},$$
$$\forall j = 1,2, \forall n \in \{1,2,\ldots,N-1\}.$$

Here, $\beta_{i,j}$ is the expected proportion of type-j demands met directly from inventory at retailer i and, hence, $(1-\beta_{i,j})\lambda_{i,j}$ is the expected number of retailer i's type-j demands requested from other retailers. The product $\Pi_{m=1}^{n-1}(1-\beta_{o_i(m),j})$ is the probability of stockout at the retailers who are before retailer $o_i(n)$ in the sequence O_i. This product is the probability that retailers whose order is less than n are unable to satisfy retailer i's transshipment request. Similarly, $\beta_{o_i(n),j}$ is the probability of no stockouts at retailer $o_i(n)$ and with this probability, $o_i(n)$ can satisfy retailer i's transshipment request. Note that, although suppressed, the $\beta_{i,j}$ values depend on the

retailers' rationing policy parameters and, therefore, so do the expected flow rates of transshipments.

Defining $\hat{\lambda}_{i,j}$ as the total demand rate at retailer i for type-j demands after all transshipment requests, we have

$$\hat{\lambda}_{i,j} = \beta_{i,j}\lambda_{i,j} + \sum_{w|i\in O_w} \gamma_{w,j,i}. \tag{2}$$

If none of the retailers can respond to the transshipment request of retailer i for type-j demand, retailer i requests an item from the central depot. Define $\gamma_{i,j,d}$ as the expected demand rate of type-j items retailer i requests from the central depot d. We have

$$\gamma_{i,j,d} = \lambda_{i,j}(1-\beta_{i,j}) - \sum_{n=1}^{N-1} \gamma_{i,j,o_i(n)}.$$

The last equality ensures that retailer i requests transshipments for all unsatisfied demands, i.e., $\lambda_{i,j}(1-\beta_{i,j}) = \gamma_{i,j,d} + \sum_{n=1}^{N-1} \gamma_{i,j,o_i(n)}$. Note that the following equality, which ensures that all demand is distributed among the retailers and the central depot, holds:

$$\sum_{\forall i\in N}\sum_{j=1,2} \lambda_{i,j} = \sum_{\forall i\in N}\sum_{j=1,2} \left(\hat{\lambda}_{i,j} + \gamma_{i,j,d}\right).$$

As stated before, the cost of not satisfying the demand immediately from stock is actually the transshipment cost to retrieve an item from another site. Suppose that retailer i requests an item from another retailer and needs to pay a high transshipment cost. This implies that it is very expensive for retailer i not to satisfy the demand of his customer for this item.

Given that $c_{i,j,o_i(n)}$ is the unit cost to transport a type-j demand from retailer $o_i(n)$ to retailer i and $c_{i,j,d}$ is the unit cost to transport a type-j demand from the central depot to retailer i, the expected cost to meet a type-j demand at retailer i via a transshipment, $\hat{\pi}_{i,j}$, can be determined. Note that $c_{i,j,d} = \pi_{i,j}$. We have

$$\hat{\pi}_{i,j} = \frac{\sum_{n=1}^{N-1} \gamma_{i,j,o_i(n)} c_{i,j,k} + \gamma_{i,j,d} c_{i,j,d}}{\sum_{n=1}^{N-1} \gamma_{i,j,o_i(n)} + \gamma_{i,j,d}}. \tag{3}$$

The third step of the approximation procedure is to solve the single-retailer problems using $\hat{\lambda}_{i,j}$ and $\hat{\pi}_{i,j}$ instead of the original values $\lambda_{i,j}$ and $\pi_{i,j}$. Therefore, each retailer re-optimizes its policy parameters given its new demand rates and penalty costs of not satisfying the demands directly from stock.

Let $(\tilde{\mathbf{r}}, \tilde{\mathbf{k}}, \tilde{\mathbf{Q}})$ be the solution of the approximation procedure. We summarize the procedure to solve the rationing and transshipment optimization problem (1) as follows:

Step 1 Determine r_i, k_i and Q_i for $j = 1, 2$ and $\forall i \in \{1, 2, \ldots, N\}$ using the enumeration-based algorithm for SiReP proposed in Sect. 4.

Step 2 Calculate $\hat{\lambda}_{i,j}$ and $\hat{\pi}_{i,j}$ for $j = 1, 2$ and $\forall i \in \{1, 2, \ldots, N\}$ using Eqs. (2) and (3), respectively.

Step 3 Set $\lambda_{i,j}$ to $\hat{\lambda}_{i,j}$ and $\pi_{i,j}$ to $\hat{\pi}_{i,j}$ for $j = 1, 2$ and $\forall i \in \{1, 2, \ldots, N\}$ and repeat Step 1.

Step 4 Set $\tilde{r}_i$ to r_i, $\tilde{k}_i$ to k_i and $\tilde{Q}_i$ to Q_i for $j = 1, 2$ and $\forall i \in \{1, 2, \ldots, N\}$.

Adjusting the demand rates at the locations to reflect the effect of transshipments is an approach utilized by other researchers (Axsäter 2003b). In addition to the demand rates, we adjust the penalty costs. This approach enables decoupling of the single-retailer problems. Note that an important advantage of the approximation procedure is that it applies to transshipment problems in which the retailers use any inventory policy, not just (r, k, Q) policies.

6 Numerical analysis

In this section, we investigate the performances of the enumeration-based approximation for SiReP and the approximation procedure for MuReP. In addition, we seek the answers for the following questions: How does the system benefit from transshipments? How is the rationing policy of a retailer affected by transshipments? Which system parameters affect the average cost the most?

6.1 Performance of the enumeration-based approximation for SiReP

We test the performance of the enumeration-based approximation for SiReP by generating 500 random parameter combinations with $h \sim U[1, 3]$, $\pi_1 \sim U[100, 10000]$, $\pi_2 \sim U[5, 100]$, $K \in \{100, 200, 500, 1000, 1500\}$, $T = 1$, $\lambda_1 = 1$ and $\lambda_2 \sim U[1, 10]$. We choose these parameter values using the numerical setup in Melchiors et al. (2000) as a reference. For each parameter combination, we determine the optimal and approximate values of the policy parameters. Let $(\tilde{r}, \tilde{k}, \tilde{Q})$ be our approximate solution, (r^*, k^*, Q^*) be the optimal solution and $C(r, k, Q)$ be the average cost of the SiReP. As a performance measure of our algorithm we use the percentage cost deviation (error), $\%\epsilon$, calculated as

$$\%\epsilon = 100 \frac{C(\tilde{r}, \tilde{k}, \tilde{Q}) - C(r^*, k^*, Q^*)}{C(r^*, k^*, Q^*)}. \tag{4}$$

In 94% of the problem instances our enumeration-based approximation finds the same result. The average and standard deviation of the cost error for the non-optimally solved instances are $\bar{\epsilon} = 0.1066\%$ and $\sigma_\epsilon = 0.0783\%$, respectively. The maximum percentage cost error is 0.1786%. If we use the approximate policy parameters and the approximate average cost (that is, replace $C(\tilde{r}, \tilde{k}, \tilde{Q})$ with the approximate average cost function with arguments $\tilde{r}$, $\tilde{k}$ and $\tilde{Q}$ in (4)), the percentage difference with the optimal results has an average, a standard deviation and a maximum value of 0.0854, 0.1089 and 0.4545%, respectively. Hence, even if we calculate the final average cost with our approximation, the percentage error is small.

Our approximation relies on replacing the random variable representing the time between the placement of a replenishment order and the rationing of inventory, i.e., H, by its mean. The approximation removes the need for integration over this random variable in the optimization of the policy parameters. This approximation results in a reduction in computational time. The average computational time reduction per cost evaluation is 0.02 s (more than 50% of the original computational time). As explained in Sect. 4, the determination of Q^* relies on enumeration. The average cost needs to be calculated for each possible value of $Q \in \{1, 2, \cdots, Q_{\max}\}$. Let the average value of $Q_{\max}$ be $\bar{Q}_{\max}$. The overall problem MuReP with N retailers requires $2N$ optimizations: one for the problem without transshipments (to calculate the new demand rates and penalty costs) and one with transshipments. Note that this approximation affects the case with $r > k$ only. Assume that half of the retailers are of type $r > k$. Therefore, for a single stock-keeping unit (SKU), the average reduction in the computational time is $\frac{2*0.02*\bar{Q}_{\max}N}{2} = 0.02\bar{Q}_{\max}N$. For example, if $\bar{Q}_{\max} = 100$, the computational time is reduced by approximately 2 seconds per SKU. Considering that even small retailers may stock tens of thousands of SKUs, and that large retailers such as Amazon may stock millions of SKUs, this computational savings is significant.

6.2 Performance of the approximation procedure for MuReP

Before using our approximation procedure for MuReP, we first test its performance under a simpler policy. Note that our approximation procedure can be applied to transshipment problems in which the retailers use any inventory policy, not just (r, k, Q) policies.

In this section, we assume that there is a single customer type and there are no fixed ordering costs, i.e., $K_i = 0, \forall i \in N$. Hence, each retailer uses a continuous-review base-stock policy, i.e., $(s-1, s)$ policy, to replenish its inventories, with transshipments allowed among retailers as described in Sect. 3.2. We use our approximation procedure in Sect. 5 to calculate the approximate base-stock levels. We compare these with the base-stock levels and the cost obtained through an enumeration-based simulation. Assuming a very high value for the maximum base-stock level, we consider all possible base-stock-level combinations and for each combination, we simulate the system to find the average cost. Then, we pick the solution with the minimum cost as the solution. This procedure is quite time-consuming. If we were to use enumeration for our original policy, i.e., the (r, k, Q) policy, we would need to enumerate $3|N|$ parameters, which would be prohibitively slow.

The replenishment lead times are constant. The demand at each retailer has a Poisson distribution, and unsatisfied customer demands are lost. For a given base-stock level, s, Karush (1957) obtains the following expressions for the stockout probability SP (s) and the average inventory level $I(s)$:

$$\mathrm{SP}(s) = \frac{f(s)}{F(s)}, \quad I(s) = \sum_{j=0}^{s}(s-j)\frac{f(j)}{F(s)} = s - \lambda T + \lambda T(\mathrm{SP}(s)),$$

Table 2 Parameter values for the experiment testing performance of MuReP procedure

Parameter	Values
N	3
T	1
λ	1; 2; 3
h_i	$\sim U[0.5, 2.5]$
$c_{i,j,k}$	$\sim U[5, 30]$
$c_{i,j,d}$	$\sim U[30, 60]$
$\text{Seq} = \begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix}$	$\begin{pmatrix} 2 & 3 \\ 3 & 1 \\ 1 & 2 \end{pmatrix}; \begin{pmatrix} 2 & 3 \\ 1 & 3 \\ 1 & 2 \end{pmatrix}$

where f denotes the Poisson probability mass function with mean λT and F the corresponding cumulative distribution function. For a given unit lost sales cost c and unit holding cost per unit time h, the average cost can be calculated by $C(s) = c\lambda \text{SP}(s) + hI(s)$ and it is straightforward to determine the optimal base-stock level.

Given this solution to the single-retailer problem, we introduce transshipments and use our approximation procedure in Sect. 5 to solve the multi-retailer problem. Table 2 summarizes the data we use. Note that the retailer sequences given in the table are not based on transshipment costs, but are simply given.

We tested 540 instances. The average percentage cost error of our procedure is 6.63% with standard deviation 9.21%. Since the simulation results rely on enumeration, we have chosen system parameters (in Table 2) that generate small base-stock levels. Therefore, even a difference of 1 in the base-stock levels can translate to a large percentage cost difference. We would expect the differences to be smaller for more realistic system parameters.

6.3 The effect of transshipments on the rationing policies

In this section, we report the results of our numerical analysis for the rationing problem with transshipments. In Sect. 6.3.1, we provide a detailed example that shows the effect of transshipments, and in Sect. 6.3.2, we perform an extensive numerical analysis to show how transshipments change the rationing policy type.

6.3.1 Base case

In their numerical analysis, Melchiors et al. (2000) consider two base cases and they call these Example 1 and Example 2. In this section, we study a system with two retailers and we use the data in Example 1 and Example 2 as the parameters for retailer 1 and retailer 2, respectively. The data are in Table 3.

Based on the definition of unit transshipment costs from the central depot, we have $c_{1,1,d} = 1000, c_{1,2,d} = 10, c_{2,1,d} = 500$ and $c_{2,2,d} = 6$. In addition, we set the unit transshipment costs between retailer 1 and retailer 2 as $c_{1,1,2} = c_{2,1,1} = 300$ and $c_{1,2,2} = c_{2,2,1} = 3$. Note that transshipments from retailers are cheaper than transshipments from the central depot. Using these parameters, we solve for the policy

Table 3 Parameter values for retailer 1 and retailer 2 (taken from Melchiors et al. 2000)

Retailer	T_i	h_i	K_i	$(\lambda_{i,1}, \lambda_{i,2})$	$(\pi_{i,1}, \pi_{i,2}$
$i = 1$	1	1	100	(1,10)	(1000,10)
$i = 2$	1	2	200	(1,5)	(500,6)

Table 4 Results for the base case

Results	No transshipments	Transshipments
$(\tilde{r_1}, \tilde{k_1}, \tilde{Q_1})$	(14, 2, 49)	(17, 1, 56)
$(\tilde{r_2}, \tilde{k_2}, \tilde{Q_2})$	(3, 12, 28)	(2, 15, 17)
$(C_1, C_2, C_{\text{total}})$	(52.48, 60.76, 113.24)	(59.41, 36.81, 96.22)

parameters of retailers when transshipments are not allowed and when transshipments are allowed. The results are in Table 4.

According to these results, before allowing transshipments, for retailer 1 it is better to place orders before starting to reject low-priority customers, i.e., retailer 1 is $k < r$ type. On the other hand, for retailer 2 it is better to start rejecting low-priority customers before placing a replenishment order, i.e., retailer 2 is $k \geq r$ type. Hence, when we introduce transshipments to this system, we expect some of retailer 2's low-priority customers' demands to be directed to retailer 1. The results in Table 4 suggest that, when transshipments are allowed, retailer 1 places more frequent orders of higher quantity, while retailer 2 places less frequent orders of lower quantity. In addition, retailer 2 employs an even harsher rationing policy, i.e., starts rejecting low-priority customer demands earlier. Transshipments result in a cost increase at retailer 1 by 13.21%, while the average cost decreases at retailer 2 by 39.41%. The total cost reduction is 15.02%. Each cost change is calculated as $\text{CR} = \frac{C^0 - C^{\text{Tr}}}{C^0}$, where C_0 is the cost before transshipments and C^{Tr} is the cost after transshipments.

Our approximation procedure relies on updating the demand rates. After transshipments we obtain $\hat{\lambda}_{1,1} = 1.0063$, $\hat{\lambda}_{1,2} = 14.4385$, $\hat{\lambda}_{2,1} = 0.9936$ and $\hat{\lambda}_{2,2} = 0.1303$. Observe that retailer 1's demand rates increase for both types of customers and especially for low-priority customers. On the other hand, retailer 2's demand rates decrease for both types of customers and especially for low-priority customers. It is easy to explain the logic behind the redistribution of the demand rates for low-priority customers. Remember that according to the rationing policies of the two retailers, retailer 2 is more likely to reject low-priority demands than retailer 1 is. Hence, it places transshipment requests with retailer 1. The two main reasons to have an increase in the high-priority demand rate of retailer 1, from 1 to 1.0063, are that retailer 1 keeps more inventory than retailer 2 and that retailer 2's unit lost sales cost for high-priority demands is half the corresponding cost for retailer 1.

The unit costs of not satisfying the demands directly from stock are $\hat{\pi}_{1,1} = 304.69$, $\hat{\pi}_{1,2} = 9.36$, $\hat{\pi}_{2,1} = 300.38$ and $\hat{\pi}_{2,2} = 3.05$. These costs are much lower compared to the original costs in Table 3. The decrease in these costs explains why after allowing transshipments both retailers decrease the portion of demand satisfied directly from stock.

6.3.2 Changes in the rationing policies

According to the example in the previous section transshipments allow retailers to employ rationing policies that suggest rejecting more low-priority customers than before. Although it is not the case in the example, it might be better for a retailer following a $k < r$-type rationing policy without transshipments to change to a $k \geq r$-type policy after transshipments are allowed. In order to test the correctness of this claim, we perform an extended numerical analysis. We use the data in Table 5.

When there are two retailers, there are 4 possible policy combinations: (i) $k_1 < r_1, k_2 < r_2$, (ii) $k_1 < r_1, k_2 \geq r_2$, (iii) $k_1 \geq r_1, k_2 < r_2$ and (iv) $k_1 \geq r_1, k_2 \geq r_2$. The data in Table 5 are constructed so that there are at least 200 instances for each pre-transshipment policy combination. We test 1000 instances in total and check how the policy combinations change once transshipments are introduced. In Table 6, we report the percentage of times a starting policy combination remains unchanged and changes to another one. In addition, we report the average cost reductions.

According to the results in Table 6, when transshipments are allowed between retailers, retailers start to incline toward a policy of type $k \geq r$, since now they have an option of satisfying their low-priority customers' demands from the other retailer for a relatively cheaper price. Retailers start rejecting low-priority customers even earlier by changing the policy from $k < r$ type to $k \geq r$ type. Hence, transshipments enable having more aggressive rationing policies and cost reductions at the same time. The greatest cost reductions occur when retailers are policy-wise similar before and after transshipments. Even if neither of the retailers' policies changes, the system benefits from transshipments.

6.4 Sensitivity analysis

The results in the previous section suggest that transshipments help in reducing the total system cost. Next, we perform sensitivity analysis to understand how the system parameters affect the percentage cost reduction. In this section, we use the data in Table 3 and change one parameter at a time to observe its effect.

Table 5 Data for extended numerical analysis—MuReP

Parameter	Value
$(\lambda_{1,1}, \lambda_{1,2})$	$(1, 1)$
$(\lambda_{2,1}, \lambda_{2,2})$	$(\sim U[1, 10], \sim U[1, 10])$
(h_1, h_2)	$(\sim U[1, 2], \sim U[1, 2])$
$(c_{1,1,d}, c_{1,2,d})$	$(\sim U[500, 1500], \sim U[5, 10])$
$(c_{2,1,d}, c_{2,2,d})$	$(\sim U[500, 1500], \sim U[5, 10])$
$(c_{1,1,2} = c_{2,1,1}, c_{1,2,2} = c_{2,2,2})$	$(\sim U[200, 400], \sim U[1, 5])$
(K_1, K_2)	$(\sim U[50, 250], \sim U[50, 250])$
(T_1, T_2)	$(1, 1)$

Table 6 Results from the extensive numerical analysis—MuReP

Before trans.	After trans.	*%of cases*	Avg. cost reduction
$k_1 < r_1, k_2 < r_2$	$k_1 < r_1, k_2 < r_2$	0.40	3.71
	$k_1 < r_1, k_2 \geq r_2$	4.27	9.98
	$k_1 \geq r_1, k_2 < r_2$	5.39	14.78
	$k_1 \geq r_1, k_2 \geq r_2$	89.93	26.75
$k_1 < r_1, k_2 \geq r_2$	$k_1 < r_1, k_2 < r_2$	0.00	–
	$k_1 < r_1, k_2 \geq r_2$	72.26	14.49
	$k_1 \geq r_1, k_2 < r_2$	0.00	–
	$k_1 \geq r_1, k_2 \geq r_2$	27.74	16.14
$k_1 \geq r_1, k_2 < r_2$	$k_1 < r_1, k_2 < r_2$	0.00	–
	$k_1 < r_1, k_2 \geq r_2$	0.00	–
	$k_1 \geq r_1, k_2 < r_2$	68.47	13.70
	$k_1 \geq r_1, k_2 \geq r_2$	31.53	15.69
$k_1 \geq r_1, k_2 \geq r_2$	$k_1 < r_1, k_2 < r_2$	0.00	–
	$k_1 < r_1, k_2 \geq r_2$	0.00	–
	$k_1 \geq r_1, k_2 < r_2$	0.00	–
	$k_1 \geq r_1, k_2 \geq r_2$	100.00	23.33

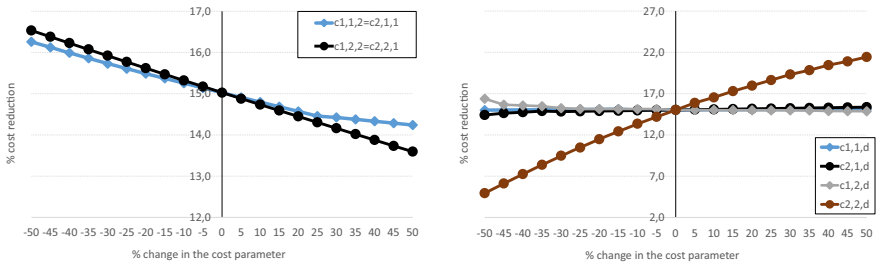


Fig. 2 **a** Inter-transshipment costs. **b** Depot transshipment costs. Effects of the unit transshipment costs on the percentage cost reduction

Intuitively, we expect transshipments to be most beneficial when it is cheap to transfer items among retailers. Figure 2a shows how $c_{1,1,2} = c_{2,1,1}$ and $c_{1,2,2} = c_{2,2,1}$ affect the percentage cost reductions achieved as a result of transshipments. When the unit cost of transshipping is low, more items are transshipped among retailers and the average cost of transshipments constitutes a significant portion of the total average cost. Hence, even a small increase in the unit cost can cause significant changes in the overall cost. In addition, our previous results suggest that most transshipments are for low-priority demands. This explains the sharper decreases in the cost reduction for low unit costs compared to high unit costs.

Figure 2b shows how the unit cost of transshipping an item from the central depot affects the cost reductions. Initially, observe that $c_{1,1,d}$, $c_{1,2,d}$ and $c_{2,1,d}$ have very

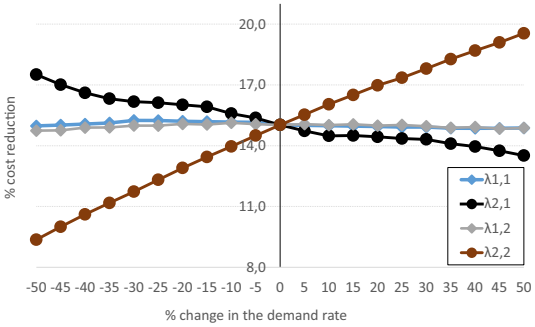


Fig. 3 Effects of the demand rates on the percentage cost reduction

insignificant effects on the cost reductions. Since retailer 1 does not use transshipments to satisfy either demand class, its transshipment costs do not contribute to cost reductions enjoyed via transshipments. Similarly, retailer 2 satisfies most of its own high-priority customer demands, so increasing $c_{2,1,d}$ does not affect cost reductions much. On the other hand, $c_{2,2,d}$ plays quite a significant role, as Fig. 2b suggests. This figure might seem counterintuitive, since having a lower unit cost results in lower percentage cost reduction, while increasing the unit cost increases the percentage cost reduction. To understand this, note that when transshipments are allowed, we require them to be made between retailers first, and then from the depot, whereas when retailer transshipments are not allowed, transshipments come only from the depot. Therefore, if the transshipment cost from the depot is low ($c_{i,j,d} < c_{i,j,w}$), the benefits of transshipments decrease, with the trend reversing as the depot transshipment cost increases.

Figure 3 shows the effect of demand rates on the percentage cost reduction achieved by allowing transshipments between retailer 1 and retailer 2. Note that the percentage cost reductions are quite insensitive to $\lambda_{1,1}$ and $\lambda_{1,2}$, the demand rates at retailer 1. Transshipments have a slight effect on retailer 1's inventory replenishment and rationing policy. Hence, any change in the demand of either type of customer does not result in change in the cost reduction, i.e., the demand rates affect the average costs before and after transshipments in the same directions and magnitude. On the other hand, retailer 2's demand rates play a more significant role in percentage cost reductions. As the rate of high-priority customer demands at retailer 2, i.e., $\lambda_{2,1}$, increases, transshipments are still beneficial, but their benefit decreases since transshipping high-priority demands is expensive. On the other hand, as the rate of low-priority customer demands at retailer 2, i.e., $\lambda_{2,2}$, increases the benefit of transshipments increases. When we compare the inventory and rationing policies of retailer 2 for low and high values of $\lambda_{2,2}$, we observe that before allowing transshipments, the main difference is in the order quantity Q_2; the higher the demand rate, the bigger the order quantity. However, retailer 2 has a high unit holding cost. Therefore, before allowing transshipments, increasing $\lambda_{2,2}$ results in higher average cost. When transshipments are allowed, retailer 2 does not need to increase its order quantity as there is another cheaper option to satisfy its low-priority demands. The average cost increases slightly as $\lambda_{2,2}$ increases, but the increase is much less than the increase in the cost before

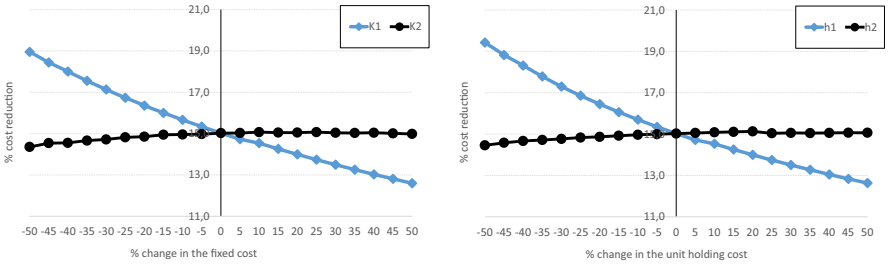


Fig. 4 **a** Fixed ordering costs. **b** Unit holding costs. Effects of the fixed ordering and the unit holding costs on the percentage cost reduction

transshipments. This explains the increase in the percentage cost reduction as $\lambda_{2,2}$ increases (Fig. 3).

Our sensitivity analysis on the effects of the fixed ordering costs K_1 and K_2 and the unit holding costs h_1 and h_2 suggests that retailer 1's cost parameters slightly affect the percentage cost reduction; the higher the K_1 and h_1, the lower the percentage cost reduction. Increasing K_1 results in less frequent orders, and increasing h_1 results in lower order quantities at retailer 1. These imply that the inventory level at retailer 1 gets lower and is not enough to satisfy all transshipment requests from retailer 2. retailer 2's cost parameters do not affect the percentage cost reduction (Fig. 4).

When the system consists of more than two retailers, each retailer chooses which retailer to request items from first. Under our objective of minimizing the total system cost, we claim that it is generally better to sort the retailers in $N\backslash i$ in increasing order of their unit transshipment cost in order to determine the sequence. In order to test this claim, we perform a numerical study using the data in Table 7.

Note that we have $c_{1,j,2} = c_{2,j,1} \leq c_{1,j,3} = c_{3,j,1} \leq c_{2,j,3} = c_{3,j,2}$ for $j = 1, 2$, which is consistent with Seq_2. For 100 random parameter combinations, we solve the problem for both sequence lists, and in all instances, we observe that Seq_2 provides lower average costs. Table 8 summarizes the percentage cost benefits achieved by using Seq_2 instead of Seq_1, i.e., $100\frac{C_1(\mathbf{r},\mathbf{k},\mathbf{Q})-C_2(\mathbf{r},\mathbf{k},\mathbf{Q})}{C_1(\mathbf{r},\mathbf{k},\mathbf{Q})}$. Here, $C_m(\mathbf{r}, \mathbf{k}, \mathbf{Q})$ is the average system cost when sequence list Seq_m is used. We calculate similar percentages for each individual retailer as well.

The only difference between Seq_1 and Seq_2 is that in Seq_1, retailer 2 places transshipment requests initially with retailer 3, although it is cheaper to transship from retailer 1. According to the results in Table 8, retailer 2 benefits the most from changing the sequence of its requests. This change benefits the overall system as well.

6.5 Inventory systems with more than three retailers

The approximation procedure for MuReP can be used for systems with any number of retailers. In this section, we summarize our results for 5-retailer and 10-retailer systems. For both systems, we consider two types of retailers. We fix T to 1. The rest of the data are in Table 9.

Table 7 Parameter values for the experiment testing the effects of transshipment sequence

Parameter	Values
N	3
T	1
$\lambda_{i,1}$	1
$\lambda_{i,2}$	$\sim U[0, 10]$
h_i	$\sim U[0.5, 1.5]$
$[c_{1,1,2} = c_{2,1,1}, c_{1,2,2} = c_{2,2,1}]$	$[\sim U[50, 100], \sim U[1, 5]]$
$[c_{1,1,3} = c_{3,1,1}, c_{1,2,3} = c_{3,2,1}]$	$[c_{1,1,2} + \sim U[0, 50], c_{1,2,2} + \sim U[0, 5]]$
$[c_{2,1,3} = c_{3,1,2}, c_{2,2,3} = c_{3,2,2}]$	$[c_{1,1,3} + \sim U[0, 50], c_{1,2,3} + \sim U[0, 5]]$
$c_{i,1,d}$	$\sim U[500, 1500]$
$c_{i,2,d}$	$\sim U[15, 20]$
K_i	200
$\text{Seq}_1 = \begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix}, \text{Seq}_2 = \begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix}$	$\begin{pmatrix} 2 & 3 \\ 3 & 1 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 3 \\ 1 & 3 \\ 1 & 2 \end{pmatrix}$

Table 8 Percentage cost benefit of Seq_2

Stats	Ret. 1	Ret. 2	Ret. 3	System
Avg.	− 0.51	22.74	0.62	8.92
SD	0.42	11.36	0.46	5.70
Max.	−0.01	51.19	2.15	25.61

Table 9 Parameter values for the retailer types

Parameter	Retailer type 1	Retailer type 2
$\lambda_{i,1}$	1	1
$\lambda_{i,2}$	$\sim U[0, 5]$	$\sim U[0, 10]$
h_i	$\sim U[2.5, 3.5]$	$\sim U[0.5, 1.5]$
$c_{i,1,w}$	$\sim U[10, 95]$	$\sim U[10, 95]$
$c_{i,2,w}$	$\sim U[0, 10]$	$\sim U[0, 10]$
$c_{i,1,d}$	$\sim U[200, 500]$	$\sim U[500, 1000]$
$c_{i,2,d}$	$\sim U[10, 15]$	$\sim U[20, 30]$
K_i	400	100

We use the enumeration-based approximation for SiReP to solve problems (without transshipments) for retailers of types 1 and 2 with 500 different parameter combinations. In all the instances, the policies for type 1 and type 2 retailers are of type $k < r$ and $k \geq r$, respectively.

Next, we allow retailers to transship and use the approximation procedure for MuReP to solve 5-retailer and 10-retailer problems. For all the instances we assume that each retailer uses a cost-ordered sequence list. For both problems, we consider multiple numbers of each retailer type. The average cost reductions are summarized

Table 10 Results for 5-retailer and 10-retailer systems

5 Retailers	(5, 0)	(4, 1)	(3, 2)	(2, 3)	(1, 4)	(0, 5)
Avg. cost reduction	28.21	29.43	30.42	30.87	33.26	41.33
10 Retailers	(10, 0)	(8, 2)	(6, 4)	(4, 6)	(2, 8)	(0, 10)
Avg. cost reduction	34.45	34.84	35.33	36.59	38.06	45.45

in Table 10. The notation (n_1, n_2) in the header rows specifies that there are n_1 and n_2 retailers of type 1 and type 2, respectively.

The results in Table 10 suggest that, independent of the number of retailers, the system enjoys significant cost reductions when transshipments among retailers are allowed. There exists a consistent cost reduction pattern when the combination of the number of type 1 and type 2 retailers changes. As the number of retailers with pre-transshipment policy of $k < r$ type increases, the percentage cost reductions increase. We observe that when transshipments are allowed, retailers start to incline toward a policy of type $k \geq r$ and the greatest cost reductions occur when more retailers are of type $k < r$ before transshipments and of type $k \geq r$ after transshipments. These results are consistent with the results in Table 6.

In today's competitive environment, most retailers do not want to lose any of their customers. On the other hand, customers benefit from having multiple competing retailers and prefer buying from the one that offers the same product for a lower price. This makes them low-priority customers. In fact, for many retailers, low-priority customers constitute a higher percentage of the total demand. This is why retailers incline toward the $k < r$-type policy. Our results indicate that by satisfying each other's customer demands through transshipments, these retailers can change their policy to the $k \geq r$ type and still enjoy significant cost benefits.

6.5.1 Alternative prioritization rule

In all the analysis and numerical results so far we assume that a type-j demand at retailer i is again a type-j demand for the retailers in the rest of the system. We claim that our approximation procedure for MuReP can be easily modified to handle any other alternative prioritization rule. In this section, we assume that all transshipment requests have low priority. This change requires modifications in the expressions introduced in Sect. 5.

$\gamma_{i,j,o_i(n)}$ is the expected demand rate of type-j items retailer i requests from retailer $o_i(n) \in O_i$. Now, since all the requests are of low priority, for $j = 1, 2$, we have

$$\gamma_{i,j,o_i(n)} = \Big((1-\beta_{i,j})\lambda_{i,j}\Big)\Big[\Pi_{m=1}^{n-1}(1-\beta_{o_i(m),2})\Big]\beta_{o_i(n),2}, \forall n \in \{1, 2, \ldots, N-1\}.$$

Here, $(1-\beta_{i,j})\lambda_{i,j}$ is the expected number of retailer i's type-j demands requested from other retailers. The product $\Pi_{m=1}^{n-1}(1-\beta_{o_i(m),2})$ is the probability that the retailers who are before retailer $o_i(n)$ in the sequence O_i reject low-priority demands. Hence, this is the probability of not satisfying retailer i's transshipment request. $\beta_{o_i(n),2}$ is

Table 11 Results for 5-retailer systems when all transshipment requests have low priority

5 Retailers	(5, 0)	(4, 1)	(3, 2)	(2, 3)	(1, 4)	(0, 5)
Avg. cost reduction	19.38	3.27	2.09	1.43	1.09	0.01

the probability that the nth retailer in the list satisfies retailer i's the transshipment request. The expressions for total demand rate at retailer i for type-j demands after all transshipment requests, $\hat{\lambda}_{i,j}$, are

$$\hat{\lambda}_{i,1} = \beta_{i,1}\lambda_{i,1},$$
$$\hat{\lambda}_{i,2} = \beta_{i,2}\lambda_{i,2} + \sum_{j=1}^{2} \sum_{w|i\in O_w} \gamma_{w,j,i}.$$

The expressions for $\gamma_{i,j,d}$ and $\hat{\pi}_{i,j}$ remain the same. Although the priority level of high-priority customers changes as retailers request transshipments, the cost of losing a customer still depends on the original priority level of the customer. This is why the expressions for $\hat{\pi}_{i,j}$ remain the same.

We compare this alternative prioritization rule with the original rule, where a retailer's high-priority demands have high priority and low-priority demands have low priority at other retailers. The results in Tables 10 and 11 are obtained using exactly the same realizations of random parameters. Here, we report our results for the systems with 5 retailers only since the results for the systems with 10 retailers are similar.

When retailers use the alternative prioritization rule to satisfy the transshipment requests the benefit of transshipments decreases. In Table 11, we report the percentage reductions in average percentage cost reductions. Hence, if $\mathrm{CR}^{\mathrm{org}}$ and $\mathrm{CR}^{\mathrm{alt}}$ represent the average percentage cost reductions due to transshipments in the original and alternative prioritization rule, respectively, the numbers in Table 11 are obtained by $100\frac{\mathrm{CR}^{\mathrm{org}}-\mathrm{CR}^{\mathrm{alt}}}{\mathrm{CR}^{\mathrm{org}}}$.

In the alternative prioritization rule all transshipment requests have low priority. This is why the customers who originally have high priority are more likely to be rejected by other retailers. Since it is costly to reject high-priority demands, the overall system cost is higher compared to the original prioritization rule. However, according to the results in Table 11, the system still enjoys significant cost reductions. The difference between the prioritization rules decreases as more retailers become $k < r$ type. Retailers who follow $k < r$-type policy have a tendency to reject a smaller number of low-priority customers. For the alternative prioritization rule, this implies fewer rejects for the customers who originally have high priority. This is why the cost benefits of the rules become similar as more retailers are $k < r$ type. Therefore, companies with multiple retailers can use the alternative prioritization rule without losing any benefit if before allowing transshipments most of the retailers are $k < r$ type.

Table 12 Results for 5-retailer systems when transshipments are allowed for high-priority demands only

5 Retailers	(5, 0)	(4, 1)	(3, 2)	(2, 3)	(1, 4)	(0, 5)
Avg. cost reduction	0.00	4.30	15.44	32.61	61.13	88.79

6.5.2 Transshipments for high-priority demands only

In all the analysis and numerical results so far we assume that the retailers can place transshipment requests for both types of customer demands. In this section, we assume transshipments for high-priority customer demands only. The rejected low-priority customer demands are satisfied directly from the central depot. We assume that a retailer's high-priority demands have high priority at other retailers.

We compare this setting with the original setting. The results in Tables 10 and 12 are obtained using exactly the same realizations of random parameters. Similar to Sect. 6.5.1, we report our results for the systems with 5 retailers only since the results for the systems with 10 retailers are similar.

When transshipments are allowed for high-priority customer demands only, the benefit of transshipments is lower compared to the original setting. In Table 12, we report the percentage reductions in average percentage cost reductions (calculated as explained in Sect. 6.5.1). According to the results, the difference between the original setting and the setting where transshipments are allowed for high-priority customer demands decreases as more retailers become $k \geq r$ type. When all retailers are $k \geq r$ type, the low-priority customer demands are rejected, i.e., satisfied from the central depot, more often anyway. This is why there is almost no difference between the original setting and the alternative setting where transshipments are allowed for high-priority customer demands only. Therefore, companies with multiple retailers can use the alternative setting without losing any benefit if before allowing transshipments most of the retailers are $k \geq r$ type.

7 Conclusions and future research directions

In this paper, we study a multi-retailer system with two types of customers. Each retailer employs a rationing, critical-level policy in the context of a continuous-review (r, Q) inventory model. We propose an accurate approximation procedure to solve the joint rationing and transshipment problem. We elaborate on how transshipments affect optimal policies of individual retailers and identify the types of systems that benefit from transshipments the most.

According to our results, transshipments reduce the penalty cost of not satisfying the customer demands directly from stock. Therefore, retailers keep less inventory and enjoy cost reductions even if they operate with lower service levels. Retailers with pre-transshipment rationing policies that suggest rejecting low-priority customer demands more aggressively, i.e., $k \geq r$, benefit from the transshipments the most. Our results indicate that when transshipments are allowed among retailers, retailers start to incline toward a policy of type $k \geq r$. Our sensitivity analysis suggests that cheap

transshipment options bring cost benefits. When the objective is cost minimization, a sequencing rule based on unit transshipment costs seems to perform better than a random rule.

We provide insights to help company managers in making decisions on how to make use of transshipments in reducing their operational costs. Depending on the retailers' pre-transshipment inventory replenishment policies, managers can design a transshipment policy for guaranteeing substantial cost reductions. More specifically, we show that as the number of retailers with pre-transshipment policy of $k < r$ type increases, the cost benefit of using transshipments increases. We also study an alternative prioritization rule, in which all transshipment requests have low priority. We conclude that the difference between the original rule and the alternative rule decreases as more retailers have a policy of $k < r$ type before transshipments. In the original setting the retailers can place transshipment requests for both types of customer demands. We alternatively analyze the setting in which transshipments are used to satisfy high-priority demands only. We conclude that the companies with multiple retailers can use the alternative setting without loss of any benefit if before allowing transshipments most of the retailers are $k \geq r$ type. Consequently, if a company has many retailers with policies of $k < r$ type before transshipments, we advise transshipments for all types of customer demands. Both prioritization rules can work. On the other hand, if a company has many retailers with policies of $k \geq r$ type before transshipments, we advise that high-priority customers are given high priority and low-priority customers are given low priority at other retailers. Transshipping only high-priority customer demands provides most of the benefit the transshipments bring, hence, can be a preferred setting.

Although our approximation procedure is robust enough to handle the case where low-priority demands of one retailer can be viewed as high-priority demands by other retailers, we do not study this in our numerical analysis. It might be interesting to see how the policy parameters change in this variation. Determination of the optimal prioritization rule is an interesting problem, too. In addition, the robustness of the procedure allows us to use it for problems with more than two customer demand classes. Here the challenge is to calculate the policy parameters for the pure rationing problem; if that can be done, we can use our algorithm to solve the problem with transshipments. This, too, can be a possible future research direction.

Our analysis is based on the assumption that the transshipping location does not hold back any inventory. An alternative pooling policy suggests holding back part of the inventory to cover future demand. The combination of rationing and this alternative policy would require two critical levels: one for the rationing policy and another for the transshipment requests. Although incorporating another decision variable makes the problem challenging, it would be interesting to investigate the interplay between these two critical levels.

References

Alvarez E, van der Heijden M, Vliegen I, Zijm W (2014) Service differentiation through selective lateral transshipments. Eur J Oper Res 237(3):824–835

Archibald T (2007) Modelling replenishment and transshipment decisions in periodic review multilocation inventory systems. J Oper Res Soc 58:948–956

Archibald T, Black D, Glazebrook K (2009) An index heuristic for transshipment decisions in multi-location inventory systems based on a pairwise decomposition. Eur J Oper Res 192:69–78

Archibald T, Sassen S, Thomas L (1997) An optimal policy for a two depot inventory problem with stock transfer. Manag Sci 43(2):173–183

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

Axsäter S (2003a) Evaluation of unidirectional lateral transshipments and substitutions in inventory systems. Eur J Oper Res 149(2):438–447

Axsäter S (2003b) A new decision rule for lateral transshipments in inventory systems. Manag Sci 49(9):1168–1179

Axsäter S (2015) Inventory control. Springer International Publishing, Switzerland

Brownell M (2013) The marketplace fairness act: why amazon wants to be taxed. http://www.dailyfinance.com/on/the-marketplace-fairness-act-why-amazon-wants-to-be-taxed/

Chew E, Lee L, Liu S (2013) Dynamic rationing and ordering policies for multiple demand classes. OR Spectr 35(1):127–151

Ching W-K, Yuen W-O, Loh AW (2003) An inventory model with returns and lateral transshipments. J Oper Res Soc 54(6):636–641

Comez-Dolgan N, Fescioglu-Unver N (2015) Managing transshipments in a multi-retailer system with approximate policies. J Oper Res Soc 66(6):947–964

Deshpande V, Cohen M, Donohue K (2003) A threshold inventory rationing policy for service-differentiated demand classes. Manag Sci 49(6):683–703

Eppen G (1979) Effects of centralization on expected costs in a multi-location newsboy problem. Manag Sci 25:498–501

Evans R (1968) Sales and restocking policies in a single item inventory system. Manag Sci 14(7):463–472

Evers P (2001) Heuristics for assessing emergency transshipments. Eur J Oper Res 129(2):311–316

Fadiloglu M, Bulut O (2010) A dynamic rationing policy for continuous-review inventory systems. Eur J Oper Res 202(3):675–685

Glazebrook K, Paterson C, Rauscher S (2015) Benefits of hybrid lateral transshipments in multi-item inventory systems under periodic replenishment. Prod Oper Manag 24(2):311–324

Grahovac J, Chakravarty A (2001) Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items. Manag Sci 47(4):579–594

Hadley G, Whitin T (1963) Analysis of inventory systems. Technical report, Prentice Hall, Englewood Cliffs, NJ

Hu J, Watson E, Schneider H (2005) Approximate solutions for multi-location inventory systems with transshipments. Int J Prod Econ 97:31–43

Hung H, Chew E, Lee L, Liu S (2012) Dynamic inventory rationing for systems with multiple demand classes and general demand processes. Int J Prod Econ 139(1):351–358

Isotupa K, Samanta S (2013) A continuous review (s, q) inventory system with priority customers and arbitrarily distributed lead times. Math Comput Model 57(5–6):1259–1269

Kaplan A (1969) Stock rationing. Manag Sci 15(5):260–267

Karush W (1957) A queueing model for an inventory problem. Oper Res 5:693–703

Kleijn M, Dekker R (1998) An overview of inventory systems with several demand classes. Technical report, Erasmus University, Rotterdam, The Netherlands

Kranenburg AA (2006) Spare parts inventory control under system availability constraints. Technical report, Eindhoven University of Technology, Eindhoven, The Netherlands

Kukreja A, Schmidt C (2005) A model for lumpy demand parts in a multi-location inventory system with transshipments. Comput Oper Res 32(8):2059–2075

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

Lee H (1987) A multi-echelon inventory model for repairable items with emergency lateral transshipments. Manag Sci 33(10):1302–1316

Liang C, Sethi S, Shi R, Zhang J (2014) Inventory sharing with transshipments: impact of demand distribution shapes and setup costs. Prod Oper Manag 23(10):1779–1794
Liu S, Song M, Tan K, Zhang C (2015) Multi-class dynamic inventory rationing with stochastic demands and backordering. Eur J Oper Res 244(1):153–163
Manjoo F (2012) I want it today. http://www.slate.com/articles/business/small_business/2012/07/amazon_same_day_delivery_how_the_e_commerce_giant_will_destroy_local_retail_.html
Melchiors P, Dekker R, Kleijn M (2000) Inventory rationing in an (s, Q) inventory model with lost sales and two demand classes. J Oper Res Soc 51:111–122
Minner S, Silver E (2005) Evaluation of two simple extreme transshipment strategies. Int J Prod Econ 93–94:1–11
Minner S, Silver E, Robb D (2003) An improved heuristic for deciding on emergency transshipments. Eur J Oper Res 148(2):384–400
Nahmias S, Demmy W (1981) Operating characteristics of an inventory system with rationing. Manag Sci 27(11):1236–1245
Needham P, Evers P (1998) The influence of individual cost factors on the use of emergency transshipments. Transp Res Part E 34(2):149–160
Olsson F (2009) Optimal policies for inventory systems with lateral transshipments. Int J Prod Econ 118(1):175–184
Olsson F (2015) Emergency lateral transshipments in a two-location system with positive transshipment leadtimes. Eur J Oper Res 242(2):424–433
Paterson C, Kiesmüller G, Teunter R, Glazebrook K (2011) Inventory models with lateral transshipments: a review. Eur J Oper Res 210(2):125–136
Paterson C, Teunter R, Glazebrook K (2012) Enhanced lateral transshipments in a multi-location inventory system. Eur J Oper Res 221(2):317–327
Seidscher A, Minner S (2013) A semi-markov decision problem for proactive and reactive transshipments between multiple warehouses. Eur J Oper Res 230(1):42–52
Seifert R, Thonemann U, Sieke M (2006) Integrating direct and indirect sales channels under decentralized decision making. Int J Prod Econ 103:209–229
Teunter R, Haneveld W (2008) Dynamic inventory rationing strategies for inventory systems with two demand classes, Poisson demand and backordering. Eur J Oper Res 190(1):156–178
Tiacci L, Saetta S (2011) Reducing the mean supply delay of spare parts using lateral transshipments policies. Int J Prod Econ 133(1):182–191
Topkis D (1968) Optimal ordering and rationing policies in a non-stationary dynamic inventory model with n demand classes. Manag Sci 15(3):160–176
Torabi S, Hassini E, Jeihoonian M (2015) Fulfillment source allocation, inventory transshipment, and customer order transfer in e-tailing. Transp Res Part E Logist Transp Rev 79:128–144
van Wijk A, Adan I, van Houtum G (2012) Approximate evaluation of multi-location inventory models with lateral transshipments and hold back levels. Eur J Oper Res 218(3):624–635
Veinott A (1965) Optimal policy in a dynamic, single product, non-stationary inventory model with several demand classes. Oper Res 13(5):761–778
Wang D, Tang O (2014) Dynamic inventory rationing with mixed backorders and lost sales. Int J Prod Econ 149(3):56–67
Wang D, Tang O, Huo J (2013) A heuristic for rationing inventory in two demand classes with backlog costs and a service level constraint. Comput Oper Res 40(12):2826–2835
Wang D, Tang O, Zhang L (2015) A note on the rationing policies of multiple demand classes with lost sales. Int J Prod Econ 165:145–154
Xu K, Evers P, Fu M (2003) Estimating customer service in a two-location continuous review inventory model with emergency transshipments. Eur J Oper Res 145(3):569–584
Yang G, Dekker R, Gabor A, Axsäter S (2013) Service parts inventory control with lateral transshipment and pipeline stock flexibility. Int J Prod Econ 142:278–289
Yang S, Liao Y, Shi C, Gao C (2014) Heuristics for solving an internet retailer's dynamic transshipment problem. Expert Syst Appl 41:5382–5389
Zhao F, Wu D, Liang L, Dolgui A (2016) Lateral inventory transshipment problem in online-to-offline supply chain. Int J Prod Res 54(7):1951–1963
Zhao H, Deshpande V, Ryan J (2006) Emergency transshipment in decentralized dealer networks: when to send and accept transshipment requests. Nav Res Logist 53:547–567