Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Inventory Control in a Spare Parts Distribution System with Emergency Stocks and Pipeline Information

Christian Howard, Johan Marklund, Tarkan Tan, Ingrid Reijnen

To cite this article:
Christian Howard, Johan Marklund, Tarkan Tan, Ingrid Reijnen (2015) Inventory Control in a Spare Parts Distribution System with Emergency Stocks and Pipeline Information. Manufacturing & Service Operations Management 17(2):142-156. http://dx.doi.org/10.1287/msom.2014.0508

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 17, No. 2, Spring 2015, pp. 142–156
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2014.0508

Inventory Control in a Spare Parts Distribution System with Emergency Stocks and Pipeline Information

Christian Howard
Traffic and Road Users, The Swedish National Road and Transport Research Institute (VTI), SE-581 95 Linköping, Sweden, christian.howard@vti.se

Johan Marklund
Industrial Management and Logistics, Lund University, 221 00 Lund, Sweden, johan.marklund@iml.lth.se

Tarkan Tan
Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands, t.tan@tue.nl

Ingrid Reijnen
Gordian Logistic Experts, 3528 BG Utrecht, The Netherlands, ingrid@rajko.nl

Motivated by collaboration with a global spare parts service provider, we consider a two-echelon inventory system with multiple local warehouses, a so-called support warehouse, and a central warehouse with ample capacity. In case of stock-outs, the local warehouses can receive emergency shipments from the support warehouse or the central warehouse at an extra cost. Our focus is on using information on orders in the replenishment pipeline, i.e., pipeline information, to achieve cost-efficient policies for requesting emergency shipments. We introduce a policy where the request for an emergency shipment is based on the time until an outstanding order will reach the stock point considered. The goal is to determine how long one should wait for stock in the replenishment pipeline before requesting an emergency shipment, and the cost effects of using pipeline information in this manner. The analysis utilizes results from queuing theory and provides a decomposition technique for optimizing the policy parameters that reduces the complex multiechelon problem to more manageable single-echelon problems. The performance of our policy indicates that there can be a significant benefit in using pipeline information.

Keywords: emergency shipments; inventory; multiechelon; pipeline information; spare parts
History: Received: September 29, 2010; accepted: October 10, 2014. Published online in *Articles in Advance* February 19, 2015.

1. Introduction

The research presented in this paper is motivated by collaboration with Volvo Parts Corporation, a global spare parts service provider with headquarters in Sweden. Volvo Parts is the supplier of aftermarket services for the Volvo Group, supporting the following business areas: Volvo Trucks, Mack, Renault Trucks, Volvo Buses, Volvo Heavy Machinery, and Volvo Penta. It follows that a core operational area for Volvo Parts is stock keeping and distribution of spare parts. These spare parts are distributed through central warehouses positioned around the world, each one responsible for serving several local markets. In each local market they have a number of local warehouses (dealers/retailers) that, in turn, serve the end customer. This includes both service and repairs of the customers' vehicles, as well as direct "over-the-counter" sales of the spare parts. The local warehouses replenish their stock by placing orders with the central warehouse.

One of the main challenges faced by Volvo Parts is how to cost effectively achieve high availability of low-demand spare parts. Looking at the central warehouse in the European market, which is responsible for supplying just over 900 stock points, 98% of the different articles ordered have a yearly demand lower than 10,000 units. This corresponds to an average customer demand of less than one unit per month at each stock point. At the same time, these articles are generally quite expensive and account for 72% of the yearly value (calculated as yearly demand multiplied by item value). To achieve high availability for these articles Volvo Parts uses a separate emergency shipment system in addition to the regular replenishment system. In most local markets they have an additional stock point referred to as a support warehouse. The support warehouse is also replenished by the central warehouse and its purpose is to provide emergency shipments to the local warehouses in cases of stock-outs. Because the support warehouse is situated

closer to the local warehouses than the central warehouse, the shipment times are much shorter than for regular replenishments (typically overnight). However, because of the extra transportation and handling activities involved, they come at a higher cost. As a last resort, the central warehouse can also provide an emergency shipment directly to the local warehouse. This type of system structure is by no means unique to Volvo Parts. It is, for instance, also used by several of their competitors.

In the organizational structure of Volvo Parts' distribution system, some of the local warehouses are owned by Volvo Parts, whereas others are independent privately owned companies. In terms of inventory control, Volvo Parts has a vendor-managed inventory (VMI) contract with the local warehouses, which gives them the mandate to control the local warehouses' inventories through a centralized IT system. More precisely, they are authorized to determine parameters (e.g., base-stock levels) in the control policies used. Apart from this, the managerial control of the local warehouses' operations is decentralized. The central warehouses and support warehouses, on the other hand, are owned and operated by Volvo Parts, which gives them full control over all emergency shipment operations. A motivation for choosing such a distribution structure is to avoid complicated incentive issues between local warehouses, typically encountered in systems where stock is shared among independently owned companies, e.g., through lateral transshipments.

Because emergency shipments are more costly than regular supply, the policy for requesting such shipments is of crucial importance. With the recent advances in information technology it is now possible to have detailed information on the state of an inventory system, such as the current positions of outstanding orders. A critical question is, how can this information be utilized in the design of emergency shipment policies? Currently at Volvo Parts, the general policy is to ask for an emergency shipment whenever a stock-out occurs. However, as recognized by the company, this is not necessarily the best strategy in terms of cost and service effectiveness. For some spare parts it might be better to back-order the demand at the given stock point, in anticipation of the next incoming regular order, instead of requesting an emergency shipment. For example, if there is a regular replenishment order in the pipeline from the central warehouse in close proximity to the local warehouse, not only will the customer waiting time be less for this incoming order, but at the same time, Volvo Parts will avoid the extra cost associated with an emergency shipment. Following this line of reasoning, it is clear that, if incoming orders are ignored, using emergency shipments can actually lead to less effective systems, compared to using only regular supply. This highlights the need for a replenishment policy that is more flexible regarding the use of emergency shipments and that utilizes information about when outstanding orders in the replenishment pipeline from the central warehouse will be arriving (which is referred to as using pipeline information).

In this paper, we focus on a single local market and introduce what we refer to as an (S, T) policy at each individual stock point. The policy parameter S is the base-stock level and T is the threshold time for back-ordering instead of emergency ordering. This threshold time is an internal decision variable that can be set individually for each product and stock point in the system and is designed to incorporate the possibility of waiting for incoming replenishment orders. In essence, T specifies how far up the pipeline to look when deciding how to replenish when there is a stock-out. When demand occurs at a specific local warehouse and that warehouse is out of stock, the demand is back-ordered if there is a regular replenishment order arriving within the set threshold time. If there is no regular order close enough in the pipeline (i.e., arriving within T time units), an emergency shipment is requested. The request first goes to the support warehouse, which will meet the demand and send an emergency shipment if there is stock on hand or stock arriving within its own threshold time. If this is not the case, the local warehouse requests an emergency shipment from the central warehouse instead. We assume that the central warehouse always can deliver and, therefore, it can be viewed as an external supplier.

Our model assumes Poisson-distributed customer demand and one-for-one ordering at all stock points, which is reasonable for the slow-moving items in Volvo Parts' product assortment. Moreover, we consider a customer waiting cost per unit and time unit, which is based on contractual obligations, downtime costs, and loss of goodwill. Given this waiting cost, along with holding costs and emergency shipment costs, we provide a method for determining base-stock levels and threshold times such that the expected system costs are minimized. Note that the determination of suitable threshold times is within the control of the company and is not something to be negotiated with the end customer. That is, the local warehouses have an obligation to provide high-quality service to the end customer (the lack of which is quantified by the waiting cost), but emergency shipment is a service provided to the local warehouses, and the threshold times are internal decision variables that are not visible to the end customer.

One of the main advantages of our (S, T) policy is its generality: when all threshold times are set to zero, this corresponds to the current situation at Volvo

Parts, where emergency shipments are used whenever a stock-out occurs. Conversely, if all threshold times are equal to the associated replenishment lead times, all demand will wait for regular replenishment, and emergency shipments are never used. Our policy therefore provides a cost performance guarantee over simple policies such as (i) always requesting emergency orders when stock-outs occur, (ii) never using this option, or (iii) always requesting an emergency order if this makes the item available faster than waiting for the next regular order to arrive. Furthermore, optimizing the threshold times provides structural results on suitable system configurations for different products (e.g., for which products the support warehouse should be used at all). This is another issue of key interest for Volvo Parts.

The contributions of our work can be summarized as follows: We address the frequently observed, yet unsolved problem of when to place emergency orders in supply chains with multiple local warehouses (or dealers, retailers, repair shops, etc.) and an organizational structure that excludes the possibility of local inventory pooling (e.g., lateral transshipments between all stock points). The importance of the problem is supported by data from Volvo Parts. We develop a multiechelon model to analyze the problem where one building block is an exact method for a local warehouse in isolation. We also provide a simple heuristic for determining S_j and T_j values for the multiechelon system, which proves to be very accurate (average relative cost increase of 0.06% in our study) and reasonably fast (average solution time of 60 seconds). Making use of data from Volvo Parts, and the flexibility of our new (S, T) policy, we obtain managerial insights by evaluating different strategies for requesting emergency shipments. Perhaps the most notable insight is that using pipeline information in a simple intuitive way can result in poor cost performance. For example, in our numerical study, choosing the quickest replenishment option (emergency stock versus pipeline stock) resulted in an average penalty of 7% (maximum of 91%), compared to optimizing the system parameters in an integrated manner using our (S, T) policy.

The remainder of this paper is organized as follows: §2 provides a review of related literature. Section 3 presents the considered model in detail and discusses the assumptions made. Section 4 analyzes a single local warehouse in isolation and provides an exact method for cost evaluation and optimization of this single-echelon system. Based on these results, a multiechelon model for the distribution system is presented together with an accurate heuristic for setting base-stock levels and threshold times for all local warehouses and the support warehouse. Section 5 evaluates the performance of the proposed heuristic and provides managerial insights regarding the value of using pipeline information. Section 6 concludes.

2. Literature Review

Our work is mainly related to the literature on lateral transshipments, dual supply, partial back-ordering, and multiechelon systems.

The lateral transshipment literature focuses on models where locations in the same echelon can share inventory by transferring items between the locations. Recent overviews of this literature are provided in Paterson et al. (2011) and Wong et al. (2006). We mention the papers by Kranenburg and van Houtum (2009) and Reijnen et al. (2009) particularly because, similar to the considered support warehouse, only a subset of local warehouses can supply transshipments. Kranenburg and van Houtum (2009) consider a two-level structure where lateral transshipments can only be supplied by the upper level. Reijnen et al. (2009) consider a structure where local warehouses can only receive a lateral transshipment if the local warehouse can be reached within a predefined time limit. Both papers assume Poisson demand, service constraints as opposed to waiting costs, and exponentially distributed replenishment lead times. The lead time assumption excludes the possibility of keeping track of outstanding orders, because of the memoryless property. Hence, although the authors of both papers recognize that the lateral transshipment time may be nonnegligible, they do not consider waiting for a replenishment order in the pipeline as an alternative to sending an emergency shipment.

A single-echelon lateral transshipment model that incorporates the option of waiting for incoming orders is provided by Yang et al. (2013). They consider a structure similar to Reijnen et al. (2009), wherein it is assumed that customers are willing to wait at a local warehouse for a given amount of time. In Yang et al. (2013), a local warehouse will wait for incoming orders, instead of requesting a lateral transshipment, if the order will arrive within this given time limit. This is similar to our assumptions but there are important differences. First, they regard the customer time limit as a given parameter and assume that the customer is satisfied if the item is received within this time (which makes it similar to a service constraint). Although our threshold times can be used in the same way (by letting them be equal to the customer time limit), in our model we consider a customer waiting cost per time unit at each local warehouse, and regard the threshold time as a decision variable. Second, they restrict all local warehouses to the same time limit, whereas we allow for different threshold times at different locations. Last, in their work they assume that demand is back-ordered

if no lateral transshipment can reach the local warehouse in time, whereas we consider an emergency shipment from the central warehouse as a last option.

Our work is also related to the literature on unidirectional lateral transshipment models (see e.g., Axsäter 2003b, Olsson 2010) because the emergency shipments occur exclusively in one direction (from the support warehouse to the local warehouses). Another paper that is related to our work is by Axsäter et al. (2013); it also analyzes the described distribution system used by Volvo Parts. However, their work is focused on minimizing costs under fill rate service constraints, and, therefore, they do not consider the customer waiting times explicitly. Moreover, they do not consider pipeline information; they assume normally distributed demand, and their model assumes that the local warehouses always order from the support warehouse when a stock-out occurs.

The main distinguishing feature of our work compared to the lateral transshipment papers mentioned above is that we consider the inventory in the pipeline before requesting an emergency (lateral) transshipment. Furthermore, our policy guarantees better or equal cost performance compared to simple policies such as complete back-ordering (no emergency orders at all), always requesting emergency shipments when stock-outs occur, or using emergency shipments when it is faster than waiting for the next incoming regular order. None of the lateral transshipment models mentioned above provide this performance guarantee. Axsäter (2003a) does suggest a heuristic decision rule for lateral transshipments that has a similar performance guarantee. This decision rule incorporates the remaining delivery times for outstanding orders. Although we also utilize this information, our use of threshold times is quite different from Axsäter's transshipment rule. Furthermore, we place an emphasis on determining replenishment policy parameters, whereas Axsäter uses simulation for evaluation and optimization of (R, Q) policies under the given transshipment rule.

In our analysis, we start with a single local warehouse in isolation. This means that, given a stock-out situation, the decision that the local warehouse faces is either to wait for regular replenishment or to satisfy the demand by an exogenous source. The dual-supply literature studies similar types of decisions, where a common assumption is that a single warehouse can choose between a regular supplier and a quicker, more expensive emergency supplier. The main difference is that, from a single-echelon modeling perspective, our local warehouse does not receive any orders from the emergency supplier. These orders are viewed as lost sales (transferred to the support warehouse in the multiechelon model). The dual-supply literature is extensive; for general overviews, we refer to Minner (2003) and Veeraraghavan and Scheller-Wolf (2008). An important work is that of Whittemore and Saunders (1977), who study a multiperiod problem and two suppliers. They show that when the lead times differ by more than one period between the two suppliers, the optimal policy is complex and highly state dependent. Moinzadeh and Nahmias (1988) use an approximation for determining parameters for an (R, Q) policy under the assumption that there can be at most one regular and one emergency order outstanding. However, they do not benchmark the cost performance of their policy against other policies.

A single-echelon dual-supply model that is closely related to our local warehouse in terms of modeling assumptions is that of Moinzadeh and Schmidt (1991). They consider an $(S-1, S)$ inventory system with Poisson demand and two exogenous suppliers. When a demand occurs, the decision to order from the more expensive emergency supplier with a shorter lead time, τ, is based on comparing the remaining times of unreserved outstanding orders to τ. Song and Zipkin (2009) reinterpret the ordering policy considered by Moinzadeh and Schmidt (1991) and show that it is equivalent to basing the ordering decision on two separate base-stock levels, S_1 and S_2 (often referred to as a dual-index policy). More precisely, an emergency order is placed when the inventory position for the downstream part of the pipeline, including stock on hand minus back orders plus outstanding order arriving to the stock point within τ time units, drops below the base-stock level S_2. Otherwise, a normal replenishment order (with lead time $L \geq \tau$) is placed to maintain the regular inventory position, including all outstanding orders, at the base stock-level S_1. This (S_1, S_2) policy is more general than the (S, T) policy in the sense that it allows for proactive emergency orders (if $S_2 > 0$), while in the (S, T) policy emergency shipments can only be requested when a shortage has occurred. On the other hand, the (S_1, S_2) policy is more restrictive than the (S, T) policy in the sense that the emergency ordering decision is based on the downstream pipeline information associated with the given emergency lead time τ, whereas in the (S, T) policy the downstream pipeline information to consider is determined by the threshold time, T, which is a control variable to be optimized. Another important difference is that the (S_1, S_2) model is restricted to either complete back-ordering or complete lost sales, whereas our (S, T) model allows for partial back-ordering (that is, a demand can be either back-ordered or lost depending on the system state at the time of a demand arrival). The partial back-ordering feature is crucial for the tractability of our proposed multiechelon model. The relationship to the (S_1, S_2) policy is further analyzed in §4.1, where we build on the results of Moinzadeh and Schmidt (1991) and Song

and Zipkin (2009) to derive the performance measures for our (S, T) policy of a local warehouse.

A single-echelon dual-supply model that focuses on the value of information, albeit under rather different modeling assumptions than those in our present work, is that of Gaukler et al. (2008). They study the value of emergency ordering based on order progression information, where outstanding orders pass through N stages. They consider an (R, Q) policy with the option of also placing emergency orders. A simulation study with heuristically determined control parameters suggests that utilizing additional order information adds significant value to emergency shipment policies, which is concurrent with our findings. Axsäter (2007) applies the same technique used in Axsäter (2003a) to derive a decision rule for when to request an emergency shipment. The analysis assumes an (R, Q) replenishment policy for which the optimal parameters are determined by simulation. Although based on a different technique, the decision rule shares similarities with our policy in the sense that it also guarantees no worse performance than to always or to never request emergency shipments.

The third stream of literature that is related to our work is that of partial back-ordering. In our model, a customer who waits for a regular replenishment to arrive is regarded as a back-ordered customer, whereas one who cannot be served within T is considered lost to the local warehouse. Basing the decision to back-order on a threshold time can therefore be viewed as a type of partial back-ordering. For partial back-ordering models concerning $(S-1, S)$ policies and Poisson demand we refer to Das (1977) and Moinzadeh (1989) and references therein. What sets this literature apart from our current work is the sole focus on single-echelon models and the fact that partial back-ordering is the result of a given customer behavior, i.e., that customers are willing to wait for a certain amount of time before leaving the system. Therefore, these papers do not investigate the value of being able to choose when to back-order a demand. Models that do investigate the value of partial back-ordering are given in Chu et al. (2001) and Rabinowitz et al. (1995). They study a single-echelon system under Poisson demand where customers are back-ordered when a replenishment order is close enough in the pipeline. Although they illustrate that there is a large potential in allowing for partial back-ordering, their analyses are approximate because they consider (R, Q) replenishment policies under the assumption that there can be at most one order outstanding. When allowing for back orders, this assumption will be violated. In particular, when Q is small there can be many orders outstanding at the same time. Thus, their model is not suitable for analyzing base-stock policies (which correspond to $Q = 1$), whereas we provide exact results for this case.

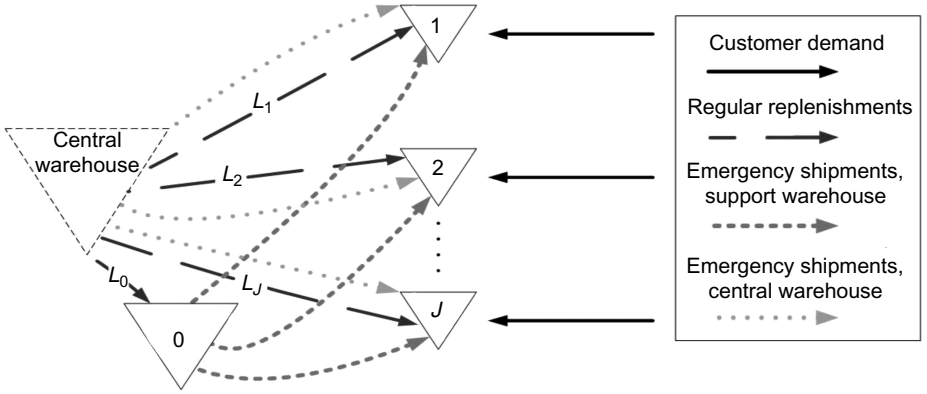
The focus on a two-level system with a single stock point in the upper echelon supplying multiple downstream facilities means that there is a relation between our work and the literature on continuous review distribution systems. For a general overview of this literature we refer to Axsäter (2003c). A more recent overview of continuous review models with one warehouse and multiple retailers is available in Axsäter and Marklund (2008). The main difference is that the upper echelon in this literature handles regular replenishment orders, whereas the support warehouse in our model handles emergency shipments. As a result, the problem formulations and solution techniques differ.

3. Problem Formulation

We consider a single-item inventory model consisting of j local warehouses (index $j \in \{1, \ldots, J\}$), a support warehouse (index $j = 0$), and a central warehouse with ample capacity (Figure 1). The rationale for the assumption of ample capacity is that the central warehouse at Volvo Parts, by a strategic decision, has very high service levels. Hence, delays due to stock-outs are rare. The support warehouse and all local warehouses apply continuous review base-stock, or $(S-1, S)$, policies, and the customer demand at local warehouse j follows a Poisson process with demand rate λ_j. These assumptions are reasonable for the slow-moving spare parts in Volvo Parts' assortment. Demand is satisfied according to a "first-come, first-served" (FCFS) rule. For fulfillment of a demand we first consider the stock on hand and the orders in the replenishment pipeline (i.e., outstanding orders en route from the central warehouse) at local warehouse j, where the demand occurred. In case local warehouse j has available stock on hand, the customer leaves directly with an item. If warehouse j is out of stock and an unreserved replenishment order will arrive within T_j time units, then the demand is back-ordered, and the customer waits until the order arrives. By "unreserved" we mean that there is no other customer demand back-ordered and waiting for the considered item.

The decision variable T_j is referred to as the threshold time at warehouse j. In case a customer demand is satisfied from stock on hand or back-ordered at local warehouse j, a new item is ordered from the central warehouse at the moment the demand occurs. The lead time L_j for these regular orders to arrive at local warehouse j is constant for a given local warehouse j, but may vary between local warehouses. If there is no stock on hand, and no stock in the replenishment pipeline that will arrive within the threshold

Figure 1 The System Under Consideration



time, the demand will be satisfied by an emergency shipment from the support warehouse or, as a last option, from the central warehouse. A demand waiting for an emergency shipment at a local warehouse is not viewed as a back order at that stock point. The reason is that the responsibility for fulfillment of that demand is now shifted to the support warehouse and the central warehouse.

When requesting an emergency shipment, the local warehouse first contacts the support warehouse, which applies the same type of (S, T) policy: (i) satisfy the demand from stock on hand by sending a shipment to the local warehouse, (ii) back-order the demand based on the pipeline inventory of reach within threshold time T_0 and send a shipment when the item arrives in stock, and (iii) deny the request for an emergency shipment. In (iii), the local warehouse requests an emergency shipment directly from the central warehouse, which can always deliver. The central warehouse emergency shipment lead time, denoted by τ_j^c, is constant for a given local warehouse, but may vary between local warehouses. The support warehouse emergency transportation time (the time from when an item leaves the support warehouse until it reaches the local warehouse), denoted by τ_j^s, is also constant. However, because stock-outs may occur at the support warehouse, the emergency shipment lead time is stochastic.

We assume a customer waiting cost b_j per unit and time unit at local warehouse j. This cost can largely be determined by estimation of the downtime costs for vehicles standing still, and penalties and discounts for delays; intangible costs, such as loss of goodwill, were also taken into consideration by a focus group at Volvo Parts. Furthermore, there is a fixed per-unit cost c_j associated with every emergency shipment from the support warehouse to local warehouse j. This cost is divided into handling and shipment costs c_j' and customer waiting costs. Analogously, there is a fixed per-unit cost p_j for every emergency shipment from the central warehouse to local warehouse j, $p_j \geq c_j$. Note that, because τ_j^s and τ_j^c are constant, the customer waiting cost for an emergency shipment in transport will be the same for all customers at local warehouse j. Thus the waiting cost, $b_j\tau_j^s$, is included in c_j (i.e., $c_j = c_j' + b_j\tau_j^s$) and, analogously, $b_j\tau_j^c$ is included in p_j. However, waiting caused by back-ordering, at a local warehouse or at the support warehouse, must be handled separately, by incurring a waiting cost that equals b_j multiplied by the duration of the back order. We also consider inventory holding costs h_j $(j = 0, \ldots, J)$ per unit and time unit for stock on hand.

Let $\mathbf{S} = (S_0, S_1, \ldots, S_J)$ be the vector of the support warehouse and local warehouse base-stock levels, and let $\mathbf{T} = (T_0, T_1, \ldots, T_J)$ be the vector of threshold times. We refer to the policy used at each stock point j as an (S_j, T_j) policy. For local warehouse j $(j = 1, \ldots, J)$ we define the following.

α_j = fraction of demand satisfied from stock on hand at local warehouse j
β_j = fraction of demand back-ordered in anticipation of pipeline stock to arrive at local warehouse j
γ_j = fraction of demand satisfied from stock on hand at the support warehouse
δ_j = fraction of demand back-ordered at the support warehouse
θ_j = fraction of demand satisfied by an emergency shipment from the central warehouse
$\psi_j = \gamma_j + \delta_j + \theta_j$, i.e., fraction of demand satisfied through emergency shipments.
EW_j = expected waiting time for an item back-ordered in anticipation of pipeline stock to arrive at local warehouse j
EV_j = expected waiting time at the support warehouse for an item requested at local warehouse j and back-ordered at the support warehouse
EIL_j^+ = expected inventory on hand at stock point j

Note that all customer demand must eventually be satisfied, i.e., $\alpha_j + \beta_j + \gamma_j + \delta_j + \theta_j = 1$. The objective is to find the **S** and **T** ($0 \le T_j \le L_j$ for all j) that minimize the expected total system cost per time unit:

$$C(\mathbf{S},\mathbf{T}) = \sum_{j=0}^{J} h_j EIL_j^+ + \sum_{j=1}^{J} b_j\beta_j\lambda_j EW_j + \sum_{j=1}^{J} c_j\gamma_j\lambda_j + \sum_{j=1}^{J} \delta_j\lambda_j(c_j + b_j EV_j) + \sum_{j=1}^{J} p_j\theta_j\lambda_j. \quad (1)$$

In (1), the first term is the expected holding costs; the second term is the expected costs for back orders at the local warehouses; the third term is the expected cost for emergency shipments sent immediately from the support warehouse; the fourth term is the expected cost for emergency shipments sent, after a delay, from the support warehouse; and, finally, the last term is the expected cost for emergency shipments sent from the central warehouse.

The settings described above match the setup of Volvo Parts' distribution system for low-demand spare parts. However, from a modeling perspective, we can instead view the request for an emergency shipment from the support warehouse as a demand lost at the local warehouse and instantly transferred to the support warehouse at the cost c_j. Analogously, an emergency shipment from the central warehouse can be viewed as demand lost for the support warehouse at an additional cost $p_j - c_j$. We therefore rearrange (1) into (2), where the first three terms are the local warehouse costs, and the last three terms are the support warehouse costs:

$$C(\mathbf{S},\mathbf{T}) = \sum_{j=1}^{J} h_j EIL_j^+ + \sum_{j=1}^{J} b_j\beta_j\lambda_j EW_j + \sum_{j=1}^{J} c_j\psi_j\lambda_j + h_0 EIL_0^+ + \sum_{j=1}^{J} b_j\delta_j\lambda_j EV_j + \sum_{j=1}^{J} (p_j - c_j)\theta_j\lambda_j. \quad (2)$$

4. Analysis

This section presents the analysis of the considered model. First, we consider a local warehouse, j, in isolation, and we show how the costs for a given (S_j, T_j) policy can be evaluated exactly, and how to optimize the decision variables. Utilizing these results, we provide an approximate cost evaluation technique for the multiechelon system and a heuristic for determining the decision variables.

4.1. Single-Echelon Model

We will from here on view the system from the modeling perspective expressed in (2). In this section we determine the first three terms in (2) exactly by analyzing a single local warehouse in isolation for a given (S, T) policy. For notational convenience we suppress the index j. Thus, demand is satisfied either directly from stock on hand, or after being back-ordered for at most T time units, or lost at a cost c per unit. Figure 2 shows a schematic representation of the local warehouse, for a given (S, T) policy.

Figure 2 Schematic Representation of a Local Warehouse

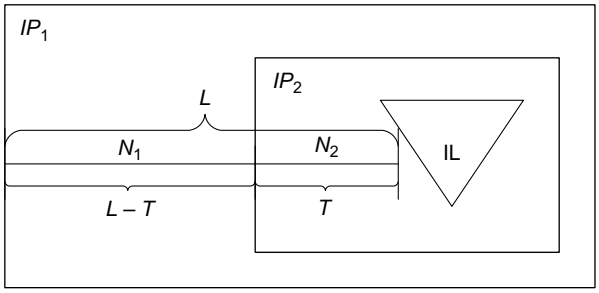


Figure 2 illustrates that the replenishment pipeline can be separated into two parts. The first part is of length $(L - T)$ time units, and an order in this part of the pipeline cannot be reserved by an arriving customer. The second part is of length T, and an (unreserved) order in this part of the pipeline can be reserved for an incoming customer demand. We define the inventory position of the whole system, IP_1, as the sum of the inventory level (IL) (defined as stock on hand minus back orders), the number of items on order in the first part of the pipeline (N_1), and the number of items on order in the second part of the pipeline (N_2). Note that, at any point in time, $IP_1 = S$ holds. Similarly, we define IP_2 as the inventory level plus the number of items in the second part of the pipeline, N_2. This implies that $IP_2 = IP_1 - N_1$ and that $0 \le IP_2 \le IP_1$. Furthermore, when $IP_2 > 0$ demand is satisfied at the local warehouse, and when $IP_2 = 0$ (or equivalently, $N_1 = S$), an arriving demand is lost for the local warehouse and satisfied by an emergency shipment. The objective is to minimize the expected costs per time unit,

$$\min C(S,T) = hEIL^+ + b\beta\lambda EW + c\psi\lambda, \quad \text{for } 0 \le T \le L,\ S \ge 0 \text{ and integer.}$$

We refer to this single-echelon model as the time-based back-ordering (TBB) model.

4.1.1. Cost Evaluation for a Given (S, T) Policy. We begin with the trivial case where $S = 0$. Because there will never be any unreserved items in the replenishment pipeline, and all demand will be lost unless $T = L$, we have

$$C(0,T) = \begin{cases} c\lambda & 0 \le T < L, \\ b\lambda L & T = L. \end{cases}$$

The result for $S = 0$ and $T = L$ (i.e., complete back-ordering) follows from Little's law. For the case where

Figure 3 The TBB Model as a Queuing Network

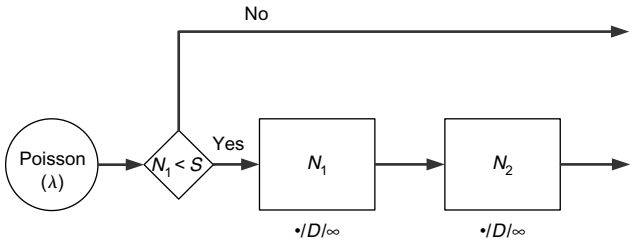
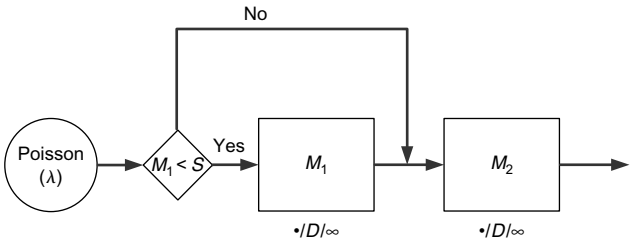


Figure 4 The DS Model as a Queuing Network



$S \geq 1$, the TBB model can be represented as a queuing network, depicted in Figure 3.

As shown in Figure 3, customers arrive at the system according to a Poisson process. A customer demand that arrives when $N_1 < S$ is either satisfied from stock on hand or back-ordered, and will therefore generate a replenishment order that goes directly into the replenishment pipeline, represented by two queuing stations. The first station, with deterministic service time $L - T$ and an infinite number of servers, represents the first part of the pipeline. The second station, with deterministic service time T and an infinite number of servers, represents the second part of the pipeline. An arriving customer that encounters S other customers in the first server, i.e., $N_1 = S$, does not generate an order and is lost to the system. Note that there are back orders in the system when $N_1 + N_2 > S$.

Analysis of this queuing network is difficult because there is no known product-form solution to the steady-state distribution of the occupancy of the system (and it is unlikely that one exists). However, as will be shown, we can circumvent this problem and obtain the performance measures we need by utilizing results from a similar queuing network, stemming from different assumptions regarding the fulfillment of customer demand. To that end, consider an alternative system where a customer facing a stock-out situation is always back-ordered (never lost) at the local warehouse, and that this always triggers a replenishment order. Furthermore, assume that the local warehouse has the option of choosing between two different suppliers, the first one with lead time L, and the second one with lead time T, $T \leq L$. The local warehouse always places its orders with the first supplier, unless a stock-out occurs and an order from the second supplier can reach the local warehouse before an (unreserved) order from the first supplier. In this case an order is placed with the second supplier. This situation corresponds to a special case of the previously mentioned (S_1, S_2) policy by Song and Zipkin (2009), where $S_2 = 0$ and $\tau = T$ (i.e., the lead time for the emergency supplier equals the threshold time). Recall that the (S_1, S_2) policy controls the ordering from each supplier by the base-stock levels S_1 and S_2 associated with the two inventory positions IP_1 and IP_2. For our purposes, the second supplier is only used when there is a stock-out and $IP_2 = S_2 = 0$. This special case, hereafter referred to as the dual-supply (DS) model, can be described by the queuing network depicted in Figure 4. In the DS model, the number of orders in each part of the pipeline is denoted M_1 and M_2, respectively.

Comparing Figure 3 with Figure 4, we see that the difference is that in the DS model customers that are blocked from entering the first station (with service time $L - T$) are not lost for the system but expedited directly to the second station (with service time T). This is sometimes referred to as "jump over blocking" in the queuing literature, and it maintains the product-form solution (see Lam 1977), making it possible to derive the steady-state distribution of the system. We will now show how this distribution can be used to obtain the performance measures for the TBB model, for given S and T.

As shown in Lam (1977), and Song and Zipkin (2009), the joint steady-state distribution of M_1 and M_2 is given by

$$\pi_{DS}(m_1, m_2) = \frac{\phi(m_1, \lambda(L-T))}{\sum_{k=0}^{S} \phi(k, \lambda(L-T))} \phi(m_2, \lambda T),$$

where $\phi(k, \mu) = e^{-\mu}\mu^k/k!$ is the Poisson probability mass function. For tractability regarding the case with $L = T$ (complete back-ordering), and the case with $T = 0$ (pure lost sales), we define $0^0 = 1$. Recall that, in order to obtain the expected costs in the TBB model, we wish to determine the expected inventory on hand (EIL^+), the fraction of demand lost for the local warehouse (ψ), and the average waiting time for a back-ordered demand (EW), in the TBB model. For the DS model we define the following.

α_{DS} = fraction of demand satisfied from stock on hand;
β_{DS} = fraction of demand back-ordered in anticipation of an item (incoming from the first supplier) already in the pipeline;
ψ_{DS} = fraction of demand back-ordered in anticipation of an incoming item from the second supplier;
EIL^+_{DS} = expected inventory on hand;

EW_{DS} = expected waiting time for a back-ordered item.

We begin with the expected inventory on hand and introduce the following lemma:

Lemma 1. *Given identical values for S, T, λ, and L, the distribution of the on-hand inventories are identical in the TBB model and the DS model.*

All proofs of lemmas, corollaries, and propositions are provided in Online Appendix A (available as supplemental material at http://dx.doi.org/10.1287/msom.2014.0508).

We note that Lemma 1 only applies to the on-hand inventories; the inventory levels in the two models are not identical. Lemma 1 implies that determining EIL^+ is straightforward using the steady-state distribution of the DS model:

Corollary 1.

$$EIL^+ = EIL^+_{DS} = \sum_{i=0}^{S}\sum_{j=0}^{S-i}(S-i-j)\pi_{DS}(i,j).$$

Next we consider the fraction of demand that is lost to the system, ψ. We can determine this fraction directly, since it is equivalent to the probability that a customer is blocked at the first station in the queuing network in Figure 3. Studying this station in isolation, it is clear that it is identical to the Erlang loss system and, hence, ψ is the Erlang loss probability, given by

$$\psi = \frac{(\lambda(L-T))^S/S!}{\sum_{k=0}^{S}(\lambda(L-T))^k/k!}.$$

The relationship between ψ and ψ_{DS} is provided in Corollary 2.

Corollary 2. *$\alpha = \alpha_{DS}$, $\beta = \beta_{DS}$, and $\psi = \psi_{DS}$.*

Corollary 2 establishes that the fractions of customers waiting in anticipation of pipeline stock are equal in the two models, and the fraction of customers lost in the TBB model is equal to the fraction of customers using the second supplier in the DS model. We will utilize this to determine the last performance measure, EW, in Lemma 2. First let EIL^-_{DS} denote the expected number of back orders in the DS model. The value of EIL^-_{DS} is obtained from the steady-state distribution as

$$EIL^-_{DS} = \sum_{i=0}^{S}\sum_{j=S-i}^{\infty}(i+j-S)\pi_{DS}(i,j).$$

Lemma 2.

$$EW = \frac{EIL^-_{DS}}{\lambda\beta} - \frac{\psi}{\beta}T.$$

The result in Lemma 2 is intuitive. Multiplying both sides with the factor $\lambda\beta$, we see that the expected number of back orders in the TBB model equals the expected number of back orders in the DS model, minus the expected number of back orders waiting for orders from the second supplier.

4.1.2. Optimization of S and T. By counterexamples it can be shown that $C(S,T)$ is neither convex in S nor convex in T, and $C(S,T)$ is not unimodal in S, implying that it is difficult to construct a simple optimization procedure. We therefore propose an enumeration procedure, where T is discretized. Using this approach one can come as close as desired to the optimal value of T by choosing a small-enough step size in the search.

Let Δ represent the step size for T. Starting at $T=0$ we increase T by Δ until we reach $T=L$. For each value of T, we start with $S=0$ and increase this variable with one unit at a time, recording the resulting total cost in each step. To find an upper bound for S (given T), we utilize that (i) EIL^+ is increasing in S, (ii) EIL^-_{DS} is nonnegative, and (iii) that the probability ψ is decreasing in S, as we know from Karush (1957). Therefore, by using Lemma 2 and rewriting the cost function as

$$\begin{aligned} C(S,T) &= hEIL^+ + b\beta\lambda EW + c\psi\lambda \\ &= hEIL^+ + b(EIL^-_{DS} - \lambda\psi T) + c\psi\lambda \\ &= hEIL^+ + bEIL^-_{DS} + \psi\lambda(c - bT), \end{aligned}$$

we conclude that, if $c \geq bT$ we can stop increasing S when $hEIL^+$ is larger than the lowest total cost found so far for the given T. Correspondingly, if $c < bT$, we stop when $hEIL^+ + \psi\lambda\,(c - bT)$ is larger than the lowest total cost found.

4.1.3. Remarks Regarding the TBB Model. In the analysis of the TBB model above, we exploit important similarities with the DS model. However, an important difference becomes apparent when applying these two models to control a given system with an emergency replenishment lead time, τ. In the DS model and in the more general (S_1, S_2) model of Song and Zipkin (2009), the inventory position IP_2 is defined by the inventory level and the outstanding orders arriving in the next τ time units. The service time of the second station (see Figure 4) must correspond to τ in order for the inventory level distributions and cost calculations in their model to hold. In the TBB model, IP_2 is defined by the inventory level and the outstanding orders arriving within the next T time units, where T is a control variable to be optimized. The impact of the emergency replenishment lead time τ is considered in the TBB model by including the associated waiting cost in the lost sales/transfer cost parameter $c = c' + b\tau$.

As mentioned earlier, the (S, T) policy does not allow for proactive emergency ordering (i.e., placing emergency orders when $IP_2 > 0$ to avoid prospective stock-outs). A generalized (S, T) policy that allows for proactive ordering can perform better, but it becomes considerably more difficult to analyze. However, we can easily obtain a lower bound for the expected costs of such a policy by removing the waiting cost component $b\tau$ in the lost sales cost c, and solving the TBB model with $c = c'$ (handling and shipment cost only). This follows as for any emergency order in the TBB model the associated waiting cost, $b\tau$, could be avoided if the order was proactively placed with the emergency supplier exactly τ time units earlier. No policy for proactive ordering can do better than this.

To assess how the TBB model performs compared to a feasible policy that allows for proactive emergency ordering, we compare it to the (S_1, S_2) model of Song and Zipkin (2009) in a numerical study detailed in Online Appendix C. The study is based on the Volvo data for the local warehouses. It shows that on average the (S, T) policy renders slightly lower costs (1.3%), but the relative performance is highly case dependent; in some instances the (S, T) policy considerably outperforms the dual-index policy and vice versa. The relative cost performance of the (S, T) policy is positively correlated with c'. When this cost component constitutes a significant part of the total per-unit emergency shipment cost c (i.e., the handling and shipment cost for emergency ordering c' is large compared to the waiting cost $b\tau$), the (S, T) policy outperforms the dual-index policy by as much as 48%. An explanation for this is that when it is more expensive to place emergency orders, it becomes more important to consider outstanding orders further up the pipeline (and allow more back-ordering) before using this expensive option. The (S, T) policy allows for this by choosing a larger threshold time T, wherease the (S_1, S_2) policy is restricted to basing the decision on the outstanding orders arriving within the emergency replenishment lead time τ. It is noteworthy that c' is incurred regardless of whether an emergency order is placed proactively or reactively. Furthermore, proactive emergency ordering tends to be less important in the Volvo settings ($S_2 = 0$ in 707 of the 840 scenarios), because the system is designed to provide short emergency shipment times ($\tau = 1$ in all cases), rendering relatively low customer waiting costs, $b\tau$.

4.2. Multiechelon Model

In §4.1 we determined the local warehouse costs exactly (i.e., the first three terms in (2)). Exact evaluation of the support warehouse costs (i.e., the last three terms in (2)) is more complicated because the demand process at the support warehouse is the sum of J "stock-out overflow" demand processes. These processes are difficult to characterize exactly, and we will therefore approximate them by independent Poisson processes. This commonly used approximation is exact in terms of the average demand rate and has been proven to work well in many situations (see, e.g., Axsäter 1990, Kranenburg and van Houtum 2009, Reijnen et al. 2009, Tiemessen et al. 2013). Moreover, our numerical tests (presented in §5.2 and Online Appendix D) indicate that the approximation works well across a wide range of problem scenarios. Note that our model is exact in the case where $S_j = 0$ and $T_j < L_j$ for $j = 1, 2, \ldots, J$, as well as the case where $T_j = L_j$ for $j = 1, 2, \ldots, J$.

Given our approximation, the demand at the support warehouse is a Poisson process with average demand rate $\lambda_0 = \sum_{j=1}^{J} \psi_j \lambda_j$. Combined with the assumption of FCFS allocation at the support warehouse, the approximation also means that all demand transferred to the support warehouse will have equal average back-order waiting times, and equal probabilities of being lost to the central warehouse. Thus, the support warehouse costs—the holding cost ($h_0 EIL_0^+$), waiting cost ($\sum b_j \delta_j \lambda_j EV_j$), and lost sales cost ($\sum (p_j - c_j)\theta_j \lambda_j$)—are determined by applying the single-echelon model to the support warehouse with demand rate λ_0, customer waiting cost b_0, and lost sales cost c_0, where

$$b_0 = \sum_{j=1}^{J} \psi_j \lambda_j b_j \Big/ \sum_{j=1}^{J} \psi_j \lambda_j \quad \text{and}$$

$$c_0 = \sum_{j=1}^{J} \psi_j \lambda_j (p_j - c_j) \Big/ \sum_{j=1}^{J} \psi_j \lambda_j$$

are the average costs for a given unit at the support warehouse.

To determine how accurate the Poisson approximation is at estimating the variance of the overflow demand, let

$N(t)$ = number of occurrences resulting from a Poisson process with rate λ on a time interval of length t, Poisson-distributed random variable with $E[N(t)] = \lambda t$, $V[N(t)] = \lambda t$;

$D_0(t)$ = demand (i.e., number of emergency shipment requests) at the support warehouse from one single local warehouse during a time interval of length t;

$X(t)$ = total accumulated time that the local warehouse is in a state such that a demand occurrence triggers a request for an emergency shipment (demand overflows to the support warehouse) on a time interval of length t, $0 \le X(t) \le t$, $E[X(t)] = \psi t$.

Because of the Poisson demand process at the local warehouse, every time the local warehouse is in a state such that demand overflows, the amount that overflows will be Poisson distributed. Furthermore, the duration of these stock-out periods at the local warehouse will be independent because of the Poisson demand and base-stock policy used. Hence, it follows that $D_0(t) = N(X(t))$. The distribution of $X(t)$ is unknown and the overflow approximation we use, denoted $\bar{D}_0(t)$, means that we replace $X(t)$ with its mean; i.e., $\bar{D}_0(t) = N(E[X(t)]) = N(\psi t)$. We introduce the following proposition:

Proposition 1. *The variance of the lead time demand at the support warehouse, $V[D_0(L_0)]$, resulting from emergency shipment requests from a single local warehouse with demand rate λ and fraction of demand satisfied by emergency shipments ψ, is bounded on the interval*:

$$V[\bar{D}_0(L_0)] \le V[D_0(L_0)] \le V[\bar{D}_0(L_0)] + \lambda^2(\psi - \psi^2)L_0^2.$$

Proposition 1 tells us that our approximation will underestimate the variance of the lead time demand, but the error is bounded, and it approaches zero as ψ approaches the extreme points of either zero or one (and is exact in the actual extreme points).

Turning to the optimization of $\mathbf{S}$ and $\mathbf{T}$, an exact procedure is difficult to obtain because the emergency shipment costs at different local warehouses are coupled to each other via the support warehouse (complete enumeration is always possible but can be very time consuming). We therefore propose a heuristic that is based on decomposing the complex multiechelon problem into simpler single-echelon problems. The heuristic is detailed in Online Appendix B.

The solution times for most problems are manageable. For instance, for the 70 problems based on data provided by Volvo Parts, presented in §5.2, the average solution time was 60 seconds (max 182 and min 18 seconds), using VBA for Excel on a laptop with a 2.4 GHz processor.

5. Numerical Experiments

The numerical experiments consist of two main studies. In §5.1 and Online Appendix D, we validate the approximations used for the multiechelon model. This is done by simulation optimization based on complete enumeration. In §5.2, we use the flexibility of the (S, T) policy to evaluate and compare different policies for requesting emergency shipments. This second study is based on data provided by Volvo Parts.

The procedure for numerical evaluation of a given problem is to use our analytical model to find near-optimal values of the policy parameters and then to use discrete-event simulation for cost evaluation. The system-optimal solution, where the S and T values are determined jointly with our analytical model, is referred to as the OPT policy, $(\mathbf{S}^{\mathrm{OPT}}, \mathbf{T}^{\mathrm{OPT}})$. The resulting expected total cost obtained by simulation is denoted $C(\mathbf{S}^{\mathrm{OPT}}, \mathbf{T}^{\mathrm{OPT}})$.

By choosing different values for the threshold time, T, the (S, T) policy can be used to implement a range of different strategies for when to request an emergency shipment. In §5.2 we evaluate four such strategies where all T_j are set to fixed values, and the S values are subsequently optimized by using our analytical model. These alternatives to the OPT policy are introduced below.

The first alternative, referred to as the "always request" (AR) policy, is based on Volvo Parts' current practice, where pipeline information is not utilized and emergency shipments are always requested when a stock-out occurs. This corresponds to the threshold times $T_j^{\mathrm{AR}} = 0$ for all j.

In the second alternative, the "never request" (NR) policy, emergency shipments are not used. The local warehouses apply complete back-ordering with threshold times $T_j^{\mathrm{NR}} = L_j$ for all j.

In the third alternative, the "quickest option" (QO) policy, a local warehouse only back-orders demand if this guarantees faster delivery than the shortest possible emergency shipment time (which occurs when there is ample stock at the support warehouse), and the support warehouse only back-orders demand if this saves time compared to an emergency shipment from the central warehouse. This translates to $T_j^{\mathrm{QO}} = \tau_j^s$ for $j = 1, \ldots, J$ and $T_0^{\mathrm{QO}} = \tau_j^c - \tau_j^s$ because all local warehouses have the same emergency shipment lead times in our study. Note that this policy is of interest only when $\tau_j^c \ge \tau_j^s$; otherwise, the support warehouse should never be used.

Although the QO policy is intuitively appealing, the threshold time T_j^{QO} is not affected by the cost parameters of the system. To that end, we propose our fourth alternative, a threshold time heuristic based on (myopically) choosing the cheapest option, referred to as the "cheapest option" (CO) policy. Assume that a stock-out occurs at local warehouse j and the closest unreserved item is δ time units away. Back-ordering the demand will result in the back-order cost $b_j\delta$, whereas requesting an emergency shipment from the support warehouse will cost c_j (ignoring that the support warehouse might also be out of stock). Hence, it is reasonable to back-order the demand if $b_j\delta < c_j$. A reasonable value for the threshold time at local warehouse j should thus be $T_j^{\mathrm{CO}} = \min(c_j/b_j, L_j)$. Following the same logic, a reasonable value for the support warehouse threshold time should be $T_0^{\mathrm{CO}} = \min((p_0 - c_0)/b_0, L_0)$, where p_0, c_0, and b_0 can, for instance, be estimated by a weighted average of each local warehouse's individual cost parameters (this is not necessary in our examples, because all local warehouses have identical values for the cost parameters).

For evaluation of the alternative policies, we used our analytical model to determine the optimal S values (denoted $\mathbf{S}^{\mathrm{AR}}$, $\mathbf{S}^{\mathrm{NR}}$, $\mathbf{S}^{\mathrm{QO}}$, and $\mathbf{S}^{\mathrm{CO}}$, respectively) given the fixed T values in each policy. The expected total cost is then determined through simulation. The measure of comparison that is used in §5.2 is the expected relative cost increase compared to the OPT policy. This measure (denoted by ΔP_{AR}, ΔP_{NR}, ΔP_{QO}, and ΔP_{CO}, respectively) is referred to as "the penalty" and is given by

$$\Delta P_{\bullet} = \frac{C(\mathbf{S}^{\bullet}, \mathbf{T}^{\bullet}) - C(\mathbf{S}^{\mathrm{OPT}}, \mathbf{T}^{\mathrm{OPT}})}{C(\mathbf{S}^{\mathrm{OPT}}, \mathbf{T}^{\mathrm{OPT}})}.$$

In Online Appendix C there are two additional studies that focus solely on the single-echelon model. The main result from the first study is that the expected cost function appears to be unimodal in T and smooth around the optimum. This suggests that the choice of T is insensitive to small errors. The second study uses data from Volvo Parts to compare the (S, T) policy with the (S_1, S_2) policy of Song and Zipkin (2009) and Moinzadeh and Schmidt (1991), as discussed earlier in §4.1.3.

5.1. Validation of the Multiechelon Model

For the multiechelon model we approximate the demand distribution at the support warehouse and we use a heuristic for determining S and T values. To provide evidence of the validity of these approximations, we compare our solutions with the ones obtained by complete enumeration using discrete-event simulation. The study focuses on how accurate our analytical model is at determining the optimal S and T values (i.e., focuses on the OPT policy), but for additional comparison we also investigate the accuracy of solely determining the S values for the AR policy used by Volvo Parts.

We consider 112 problem scenarios, evaluated both for OPT and AR, which renders a total of 224 problems. Cases assuming both identical and nonidentical local warehouses are included in the study. Our analytical model found the exact same solution as the simulation optimization in 208 out of the 224 problems. For the 16 problems where the analytical solutions differed, this resulted in an average relative total cost increase of 0.9% (maximum 1.6%), implying an average relative increase of only 0.06% over all 224 problems. Hence, the model seems to produce accurate results. Given the unsystematic occurrences of alternative solutions, there does not appear to be any clear pattern as to when the analytical model will render a different solution than the simulation optimization. Moreover, our method appears to work well in both cases of identical and nonidentical local warehouses. Further details are provided in Online Appendix D.

5.2. Results from the Case Company

This section presents a simulation study based on data from the Volvo Parts Spanish market. In total, 70 low-demand spare parts are selected for which the $(S-1, S)$ policy is an appropriate choice.

On the Spanish market, Volvo Parts has 63 local warehouses, but not all parts are sold at all locations. For the purpose of this work, we consider 12 local warehouses for each item (i.e., $N = 12$ for all problems). The demand rates at the local warehouses vary between 0.003 and 0.560 per day. The lead times for regular replenishments from the central warehouse are six days for all local warehouses. For the support warehouse, a regular replenishment from the central warehouse takes three days.

The holding costs are the same at all stock points, and all the costs have been normalized so that the holding costs are equal to one for all problems. The waiting costs are the same for all retailers, but they vary depending on item. For the items considered, the customer waiting costs per unit and day are between 17 and 625 times higher than the holding cost. These costs, that reflect the consequences of providing poor service, are based on focus group discussions at Volvo Parts. In these discussions contractual obligations, downtime costs, and loss of goodwill were taken into consideration. The main component in the emergency shipment costs is the cost of transportation, where Volvo Parts pay their transporters on a weight-per-kilometer basis. Recall that the extra costs of picking, packing, and receiving and the cost for waiting for units that have left the support warehouse (or the central warehouse) are also included in c_j (or p_j). It should be noted that, although we exclude the central warehouse from the present analysis, operations at the central warehouse have been analyzed in order to obtain accurate cost estimates. Emergency shipments that are dispatched from the support warehouse, which is situated close to Madrid, reach the local warehouses by truck within a day ($\tau_j^s = 1$). The cost of this, c_j, varies depending on item, and it is between 32 and 877 times the holding cost. The emergency shipments from the central warehouse also take one day ($\tau_j^c = 1$) but are more expensive (p_j is between 76 and 8,472 times the holding cost for the various items). The main reason for the higher costs of these shipments is that the central warehouse is situated in Ghent, in Belgium, and therefore they have to use air freight (as opposed to land freight for regular shipments) when providing emergency shipments from this location. When facing a stock-out in Volvo Parts' system, one would typically be interested in what day the desired part will arrive. It is therefore reasonable to discretize the T values to whole days.

We first compare Volvo Parts' current AR policy with the OPT policy. Detailed results are provided in

Tables E1–E3 in Online Appendix E. Note that the scenarios are sorted according to largest ΔP_{AR}, and that the local warehouses are sorted according to largest λ within each scenario. The standard deviations of all simulated values are always less than 1% of the mean. The average ΔP_{AR} over the 70 scenarios is 15.4%, with a maximum of 106% and a minimum of 2.4%. Thus, it can be very costly to ignore pipeline information and always request an emergency shipment.

In Tables E1–E3 we see that for 22 of the 70 parts, it is optimal to use the NR policy ($T_j^{OPT} = 6$ for $j = 1, \ldots, 12$). Using emergency shipments appears to be the wrong strategy for these spare parts, and it follows that the ΔP_{AR} values are among the highest in these cases. Not surprisingly, the common attribute for these parts is that the costs of emergency shipments are relatively high compared not only to the waiting but also to the holding costs. Figure 5 shows how the values of ΔP_{AR} are positively correlated with the ratio of the support warehouse emergency shipment cost to the customer waiting cost, i.e., the ratio c/b, (a similar correlation was also recorded with the ratio p/b). However, even in cases with low c/b and p/b ratios, and thus low ΔP_{AR}, Tables E1–E3 show that it is never optimal to use the current AR policy. That is, even when emergency shipments are cheap, one should not always request an emergency shipment, but in these cases most often choose the quickest option (i.e., $T_j = 1$).

Another interesting result, from Volvo Parts' perspective, is that the solutions point to a rather infrequent use of emergency shipments from the central warehouse. From Tables E1–E3 we deduce that these emergency shipments are only utilized in 20 of the scenarios considered. Furthermore, of these 20 scenarios, there are only three cases where the central warehouse is the sole provider of emergency shipments (i.e., $S_0^{OPT} = 0$ and $T_0^{OPT} = 0$ in scenarios 62, 68, and 70). However, this does not mean that emergency shipments do not bring significant value to Volvo Parts. Table 1 presents the average and maximum penalties over the 70 scenarios for all alternative emergency shipment policies. Values for ΔP_{AR}, ΔP_{NR}, ΔP_{QO}, and ΔP_{CO} for each of the 70 items are available in Tables E1–E3.

Figure 5 ΔP_{AR} **as a Function of** c/b

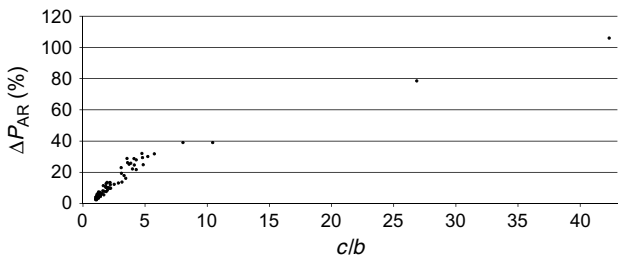


Table 1 **Average and Maximum Penalties for All Alternative Policies**

	AR	NR	QO	CO
Average ΔP (%)	15.4	6.4	7.2	0.4
Maximum ΔP (%)	106.0	49.5	90.8	4.2

We see that the penalties for the NR policy are relatively high, implying that for many items the support warehouse brings significant value to the system. Table 1 reveals that the QO policy also performs poorly in general. Hence, basing the emergency shipment decision only on the emergency shipment lead times and ignoring the cost structure of the system can be costly. We also see that the CO policy outperforms the other simpler emergency shipment policies. With a small average penalty, it appears to be an attractive heuristic that does not require enumeration of the threshold times.

Although our tests have illustrated that incorporating pipeline information in the emergency shipment decision can be very beneficial from a cost perspective, and the (S, T) policy in principle is easy to implement, it is not without challenges. A key issue, apart from the potential technical difficulties associated with integrating new decision tools in existing enterprise resource planning (ERP) systems, is that information about cost parameters and pipeline inventories may not be readily available in the firm's IT systems. With respect to the cost parameters, the emergency shipment costs (c_j and p_j) and the waiting costs (b_j) for all warehouses and items are in general the most difficult to obtain. This is, for example, the case at Volvo Parts where (like in many other companies) inventory policies have traditionally been based on minimizing holding costs at each local facility under service level constraints. Note that the "always request" (AR) policy that Volvo Parts uses, which in our study has the worst cost performance, makes perfect sense when each inventory location focuses on customer service, and costs for emergency shipments and waiting are disregarded. As for the lack of pipeline information, it is a diminishing implementation issue considering the fast development of IT and mobile technology (including radio-frequency identification (RFID) and track and trace systems). Still, if the required pipeline information is not yet available, as in the case of Volvo Parts, investments to upgrade the IT systems are required before the (S, T) policy, or for that matter, the heuristic CO or QO policies, can be implemented.

It is worth noting that, in the absence of pipeline information, an alternative use for our model could be to analyze whether the AR or NR policy performs better for each product separately. That is, to only provide emergency shipments for some spare

parts and apply complete back-ordering for the others. For our test data, this approach of considering $\min(\Delta P_{AR}, \Delta P_{NR})$ for each scenario results in an average penalty of 3.5% with a maximum penalty of 19.5%. Thus, on average the strategy may be a viable option. However, to avoid substantial penalties for some spare parts it is necessary to incorporate pipeline information. It is also worth noting in Tables E1–E3 in Online Appendix E that choosing the better of the NR and CO policies in each scenario results in a performance on par with the OPT policy for our test data.

6. Summary and Concluding Remarks

In this paper we have introduced a new policy, the (S, T) policy, for controlling an inventory system consisting of a number of local warehouses, a support warehouse, and a central warehouse with ample capacity. The policy can be used to simultaneously determine stock levels and strategies for requesting emergency shipments. We have provided an accurate heuristic for determining costs and optimizing decision variables that is based on exact evaluation of a local warehouse in isolation.

Our results indicate that significant cost benefits can be reaped by applying our new policy. The determination of suitable stock levels and the time one should wait for stock in the replenishment pipeline are complex decisions. Ignoring pipeline information and applying simple decision rules, such as always requesting a shipment or never requesting a shipment when facing a stock-out situation, can be far from cost optimal. Moreover, incorporating pipeline information in a simple way, such as only basing the decision on the emergency shipment lead times, can produce equally poor results.

Regarding implementation of our proposed policy, it is crucial that the necessary information about cost components and pipeline inventories is available in the system. With modern track and trace systems, it is possible to have real-time information on incoming orders at all warehouses. Furthermore, at many companies a warehouse knows when an order will be arriving, even without these advanced information systems, because of strict routines and high delivery reliability. This is, for example, the case at Volvo Parts, although they are also in the process of upgrading their information systems.

For future research, we believe that significant advantages can be achieved by using our (S, T) policy in more general systems where emergency shipments are utilized. One step in this direction is to extend our current model scope to include the central warehouse, and to question the high service level requirements currently used at this location. This would require that all markets served by the central warehouse are taken into consideration. Other interesting extensions would be to consider batch ordering, more general demand distributions, and proactive emergency ordering.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2014.0508.

Acknowledgments

The authors thank Volvo Parts Corporation for their collaboration and data sharing.

References

Axsäter S (1990) Modelling emergency lateral transshipments in inventory systems. *Management Sci.* 36(11):1329–1338.
Axsäter S (2003a) A new decision rule for lateral transshipments in inventory systems. *Management Sci.* 49(9):1168–1179.
Axsäter S (2003b) Evaluation of unidirectional lateral transshipments and substitutions in inventory systems. *Eur. J. Oper. Res.* 149(2):438–447.
Axsäter S (2003c) Supply chain operations: Serial and distribution inventory systems. Graves SC, de Kok T, eds. *Handbooks in Operations Research and Management Science*, Vol. 11, Supply Chain Management: Design, Coordination and Operation (North-Holland, Amsterdam), 525–559.
Axsäter S (2007) A heuristic for triggering emergency orders in an inventory system. *Eur. J. Oper. Res.* 176(2):880–891.
Axsäter S, Marklund J (2008) Optimal position based warehouse ordering in divergent two-echelon inventory systems. *Oper. Res.* 56(4):976–991.
Axsäter S, Howard C, Marklund J (2013) A distribution inventory model with transshipments from a support warehouse. *IIE Trans.* 45(3):309–322.
Chu CW, Patuwo BE, Mehrez A, Rabinowitz G (2001) A dynamic two-segment partial back-order control of (r, Q) inventory system. *Comput. Oper. Res.* 28(10):935–953.
Das C (1977) The $(S-1, S)$ inventory model under time limit on backorders. *Oper. Res.* 25(5):835–850.
Gaukler GM, Özer Ö, Hausman WH (2008) Order progress information: Improved dynamic emergency ordering policies. *Production Oper. Management* 17(6):599–613.
Karush W (1957) A queuing model for an inventory problem. *Oper. Res.* 5(5):693–703.
Kranenburg AA, van Houtum GJ (2009) A new partial pooling structure for spare parts networks. *Eur. J. Oper. Res.* 199(3):908–921.
Lam SS (1977) Queuing networks with population size constraints. *IBM J. Res. Development* 21(4):370–378.
Minner S (2003) Multiple-supplier inventory models in supply chain management: A review. *Internat. J. Production Econom.* 81–82(1):265–279.
Moinzadeh K (1989) Operating characteristics of the $(S-1, S)$ inventory system with partial backorders and constant resupply times. *Management Sci.* 35(4):472–477.
Moinzadeh K, Nahmias S (1988) A continuous review model for an inventory system with two supply modes. *Management Sci.* 34(6):761–773.
Moinzadeh K, Schmidt CP (1991) An $(S-1, S)$ inventory system with emergency orders. *Oper. Res.* 39(2):308–321.
Olsson F (2010) An inventory model with unidirectional lateral transshipments. *Eur. J. Oper. Res.* 200(3):725–732.
Paterson C, Kiesmüller G, Teunter R, Glazebrook K (2011) Inventory models with lateral transshipments: A review. *Eur. J. Oper. Res.* 210(2):125–136.
Rabinowitz G, Mehrez A, Chu CW, Patuwo BE (1995) A partial back-order control for continuous review (r, Q) inventory

system with Poisson demand and constant lead time. *Comput. Oper. Res.* 22(7):689–700.

Reijnen IC, Tan T, van Houtum GJ (2009) Inventory planning for spare parts networks. Working paper, Eindhoven University of Technology, Eindhoven, The Netherlands.

Song JS, Zipkin P (2009) Inventories with multiple supply sources and Networks of queues with overflow bypasses. *Management Sci.* 55(3):362–372.

Tiemessen HGH, Fleischmann M, van Houtum GJ, van Nunen JAEE, Pratsini E (2013) Dynamic demand fulfillment in spare parts networks with multiple customer classes. *European J. Oper. Res.* 228(2):367–380.

Veeraraghavan S, Scheller-Wolf A (2008) Now or later: A simple policy for effective dual sourcing in capacitated systems. *Oper. Res.* 56(4):850–864.

Whittemore AS, Saunders SC (1977) Optimal inventory under stochastic demand with two supply options. *SIAM J. Appl. Math.* 32(2):293–305.

Wong H, Van Houtum GJ, Cattrysse D, Van Oudheusden D (2006) Multi-item spare parts systems with lateral transshipments and waiting time constraints. *Eur. J. Oper. Res.* 171(3):1071–1093.

Yang G, Dekker R, Gabor AF, Axsäter S (2013) Service parts inventory control with lateral transshipments and pipeline stock flexibility. *Internat. J. Production Econom.* 142(2):278–289.