Manufacturing & Service Operations Management

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

Determining Optimal Parameters for Expediting Policies

Raik Özsen, Ulrich W. Thonemann

To cite this article:
Raik Özsen, Ulrich W. Thonemann (2015) Determining Optimal Parameters for Expediting Policies. Manufacturing & Service Operations Management 17(1):120-133. http://dx.doi.org/10.1287/msom.2014.0506

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 17, No. 1, Winter 2015, pp. 120–133
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs®
http://dx.doi.org/10.1287/msom.2014.0506

Determining Optimal Parameters for Expediting Policies

Raik Özsen, Ulrich W. Thonemann
Department of Supply Chain Management and Management Science, University of Cologne, 50923 Köln, Germany
{raik.oezsen@uni-koeln.de, ulrich.thonemann@uni-koeln.de}

We consider an inventory policy that expedites delivery times of open orders if the inventory level drops below a certain threshold. By expediting open orders, back orders can be reduced. Order expediting is costly, and we include various types of expediting costs in the model. We prove structural properties of the model and show how the optimal parameters of the expediting policy can be computed efficiently. The expediting policy is easy to implement and, for situations with variable expediting cost only, the structure of the policy is optimal. For situations with nonvariable expediting costs, the expediting policy that we consider is generally not optimal. The optimal policy can be computed by dynamic programming, but this approach is computationally feasible only for small-problem instances. We conduct numerical experiments that are based on data from the service division of a global equipment manufacturer to evaluate the performances of the expediting policy and the optimal policy. The results show that substantial cost savings can be achieved by order expediting and that the expediting policy realizes a great share of the cost-saving potential offered by order expediting.

Keywords: order expediting; lead time flexibility; inventory management
History: Received: July 11, 2013; accepted: September 23, 2014. Published online in *Articles in Advance* December 11, 2014.

1. Introduction

The efficient management of service parts inventories is a challenging task. Service parts inventory managers must handle a large variety of parts that are often expensive, have low demand rates and long lead times, and require high investments in inventory. While working with the service division of one of Germany's leading equipment manufacturers, we observed an order-expediting practice that can partly explain how the company achieves high availability at relatively low inventory investments. Inventory managers receive messages from the information system if the inventory level of a part drops below a certain level. An inventory manager who receives a message analyzes the on-hand inventory level, the demand forecast, and the open orders of the part and can decide to contact the supplier of the part to expedite delivery. This approach enables the company to achieve higher availability with a given inventory investment than it would achieve without order expediting. However, implementing order expediting is costly. It requires additional personnel in inventory management and increases costs at the supplier and in the logistics network.

Order expediting has not only been implemented at the equipment manufacturer but is commonly used in the service parts industry. We conducted the same survey at two business conferences[1] and asked the participants about the actions that they take if the inventory level of a part is low and they risk a back-order situation. The participants were service parts managers from the European divisions of medium- to large-sized discrete manufacturing companies. Inventory managers from 32 companies provided answers, and 28 responded that they expedite orders to avoid back orders. At 12 of the 28 companies, the expediting decision is supported by an information system. The results indicate that order expediting is commonly used, but that it is only partly supported by information systems, which is a prerequisite for implementing expediting policies efficiently. At the equipment manufacturer that motivated our research, information system support has been implemented by using an alert-level concept, wherein the inventory manager receives a message if the inventory level drops below the alert level. The alert levels are set by inventory managers using simple heuristics.

To expedite open orders optimally, the benefits of expediting must be balanced with the costs. The benefits can be quantified by the cost savings in inventory cost. Quantifying the costs of order expediting is more difficult, because the cost components are less transparent. To expedite an order, the supplier must be contacted, the supplier must make arrangements to speed up production and/or shipment of the order,

[1] The survey was conducted at the following conferences: "Optimizing the International Spare Parts Management in the Machinery Industry," Marcus Evans Conferences, January 24–25, 2013, Berlin and "Spare Parts Business Platform 2013," Copperberg, February 7–8, 2013, Stockholm.

and the delivery must be rearranged. The costs of such activities are partly fixed but can also depend on the number of units expedited and can be affected by the number of orders and batches that are expedited.

The literature on order expediting is relatively scarce. Some optimal approaches have been developed (e.g., Lawson and Porteus 2000), but the focus has been on heuristic solution approaches (e.g., Chiang 2002, Minner et al. 2003, Zhou and Chao 2010). In our base model, expediting cost depends only on the number of expedited units and the number of expedited periods. For this model, the dynamic programming approach by Lawson and Porteus (2000) can be used to obtain an optimal solution. However, more comprehensive expediting cost functions, like the one that we consider in our full model, have not been analyzed in the literature. Unlike previous research, we do not rely on dynamic programming but use a modeling approach that allows us to find the optimal solution by solving a simple convex integer optimization problem. This solution approach can be implemented easily and can solve large problems efficiently.

The main contribution of the paper is a novel approach to modeling and optimally solving an order-expediting policy that allows for quite general expediting cost functions. One of the main challenges is that the objective function is not convex in the decision variables. However, we can prove convexity of the objective function in an auxiliary decision variable, and we can derive a unique relationship between the auxiliary decision variable and the decision variables of the actual problem. Utilizing this relationship, we can restrict the optimization to a one-dimensional optimization of a convex objective function. We quantify the benefits of order expediting by conducting numerical experiments using data from the equipment manufacturer that motivated our research. The results show that the benefits of order expediting can be substantial and help us to identify situations in which it is particularly beneficial.

We consider a policy that expedites the delivery times of open orders if the inventory level drops below a certain threshold. Although this policy is easy to implement, it does not always capture the full cost-savings potential that order expediting offers. To quantify the performance gap between the expediting policy that we consider and the optimal policy, we determine the optimal solution for a set of problems by dynamic programming. For small problems that we can solve optimally, the expediting policy captures a large portion of the cost-savings potential for most problems.

2. Literature Review

Our model is closely related to the literature on inventory management with order expediting, where expediting decisions can be made during the replenishment lead time after orders have been placed. Optimal and heuristic approaches have been developed, and we discuss them next.

Optimal solution approaches exist for serial inventory models. If we interpret the time that it takes an order to move from one stage to the next stage as a lead time period, then these models can be used to analyze single-echelon structures with infinite supply, such as the one that we consider. Lawson and Porteus (2000) consider the periodic review system of Clark and Scarf (1960) and include an expediting option. The regular lead time between two consecutive stages is one period but can be reduced to an instant delivery. Lawson and Porteus (2000) assume that shipping costs are linear and additive and show that the optimal solution can be obtained by recursively solving nested single-dimensional convex optimization problems. Muharremoğlu and Tsitsiklis (2003) generalize the work of Lawson and Porteus (2000) by allowing the expediting of one unit through several stages to cost less than the sum of the one-stage expediting costs.

Models that are also related to our research are presented by Song and Zipkin (2013) and Berling and de Albeniz (2011). Both models consider supply streams, i.e., continuous-time, continuous-stage versions of supply chains. The authors consider compound Poisson demand and use solution approaches that are based on differential equations. Song and Zipkin (2013) consider a model where the delivery speed can be changed. Inventory holding cost is increasing along the supply stream, and the objective is to choose a speed that minimizes the sum of inventory holding and back-order penalty costs. Berling and de Albeniz (2011) consider a more general model that allows for a speed-dependent transportation cost.

Heuristic solution approaches have also been addressed in literature. Zhou and Chao (2010) argue that computing the optimal solution for the models presented by Lawson and Porteus (2000) is a tedious process and propose a faster heuristic solution approach that generates solutions that are close to optimal. Minner et al. (2003) also develop a heuristic solution approach, but for a divergent supply chain structure.

For inventory systems with fixed order costs, only heuristic solution approaches have been developed. Allen and D'Esopo (1968) include an expediting level concept in the classical continuous review (R, Q) policy. When the on-hand inventory reaches the expediting level, the residual lead time of the outstanding order is reduced to an expediting lead time. The authors approximate the cost function and provide a heuristic optimization approach. Chiang (2002) extends the model and introduces an additional parameter, the threshold time. An outstanding order is expedited only if the on-hand inventory reaches the expediting level before the threshold time. This approach avoids the case in which orders are expedited that would arrive in the near future without expediting. The model is solved by complete enumeration. Allen and D'Esopo (1968) and Chiang (2002) assume that at most one order is outstanding.

The dynamic programming-based approach of Lawson and Porteus (2000) and the related approaches reviewed above correspond to the base case of our model, where only variable expediting costs are considered. However, the approaches cannot be used for situations with expediting cost components other than the variable expediting cost. We contribute to the literature by developing a modeling and solution approach that is not based on dynamic programming and allows us to find solutions very efficiently.

Also related to our research is the literature on inventory management with multiple lead time options, where replenishment lead time decisions are made when an order is placed or at some prespecified time in the future. Several authors have analyzed *multisourcing models*, where faster-than-regular supply options can be used if inventory is low (e.g., Barankin 1961, Chiang and Gutierrez 1996, Tagaras and Vlachos 2001) or if a stochastic lead time exceeds a threshold (e.g., Sculli and Wu 1981, Kelle and Silver 1990, Sedarage et al. 1999). Recent contributions include Veeraraghavan and Scheller-Wolf (2008), Sheopuri et al. (2010), and Zhou et al. (2011). Veeraraghavan and Scheller-Wolf (2008) introduce the dual-index base stock policy, which uses two order-up-to levels. Sheopuri et al. (2010) generalize the class of dual-index policies studied by Veeraraghavan and Scheller-Wolf (2008). Zhou et al. (2011) derive an analytical solution for a periodic review fixed-lifetime perishable problem with two order options during a cycle.

We note that lost-sales inventory models can often be interpreted as special cases of multisource models, where one supplier instantaneously fills demands that are not satisfied by regular replenishments. For a comprehensive review of research on lost-sales inventory models, we refer the reader to Bijvank and Vis (2011).

Postponed lead times choice models are somewhat between multisourcing and order-expediting models. In these models, the lead time is chosen after the order has been placed. Chiang (2002) considers a model where the lead time consists of a manufacturing lead time and a delivery lead time. The manufacturing lead time is fixed, but there exist a regular mode and an emergency mode for the delivery lead time. The expediting decision is made at the end of the manufacturing lead time. Duran et al. (2004) consider a similar model and develop an optimal solution approach. Chiang (2010) and Jain et al. (2010) extend the model by allowing partly expedited deliveries.

Despite the similarities between inventory models with multiple lead time options and order-expediting models, there is an important difference. In inventory models with multiple lead time options, the lead time decision is made at the time when the order is placed (or at some fixed time after the order has been placed). In expediting models, first orders are placed and then the inventory system is monitored continuously or periodically to decide if orders are expedited; i.e., expediting decisions are made after orders have been placed. Inventory models with multiple lead time options and order-expediting models generally use different model structures and solution approaches.

In this paper, we extend the literature on order expediting by modeling expediting cost more generally than the existing literature and by developing an efficient solution approach.

3. Modeling Framework

We consider a periodic review base stock inventory system under an average cost criterion with stochastic demand and the option to expedite orders. We denote the demand in period t by D_t. D_t are independent and identically distributed discrete random variables. Regular replenishment orders arrive after a deterministic lead time of L periods. Expedited units arrive after a nonexpeditable lead time of $L_n < L$ periods, such that expediting is only feasible for orders placed within the previous $L_e = L - L_n$ periods. Demand that is not filled during a period is back-ordered.

We denote the expected inventory level at the end of a period by $\mathbb{E}I$, the expected back-order level of a period by $\mathbb{E}B$, the expected number of units and periods that are expedited in a period by $\mathbb{E}C^v$, the probability that expediting is used in a period by $\mathbb{E}C^f$, the expected number of batches that are expedited in a period by $\mathbb{E}C^b$, and the expected number of orders from which units are expedited in a period by $\mathbb{E}C^o$. The review period is one period.

Our performance measure is the expected cost per period:

$$Z = h \cdot \mathbb{E}I + b \cdot \mathbb{E}B + c_v \cdot \mathbb{E}C^v + c_f \cdot \mathbb{E}C^f + c_b \cdot \mathbb{E}C^b + c_o \cdot \mathbb{E}C^o. \quad (1)$$

Unit inventory holding cost is h and unit back-order penalty cost is b. Variable expediting cost is c_v, fixed expediting cost is c_f, expediting cost per batch is c_b, and expediting cost per order is c_o. Expediting cost per batch, c_b, is charged for each batch of up to q units that are expedited. The batch size q is an exogenously given parameter, and we do not optimize this parameter.

With the nonexpeditable section of the lead time and with the four types of expediting cost that we use, we obtain a quite general setup that can be parameterized to model a wide variety of situations. With the nonexpeditable section of the lead time L_n, we can model situations where a portion of the lead time cannot be expedited, for instance, because final testing must be performed on a part or a part must be shipped via ocean freight or because expediting itself takes a significant amount of time.

The variable expediting cost c_v is used to capture situations where expediting cost is proportional to the number of units that are expedited *and* to the number of periods by which a unit is expedited, i.e., where the cost of expediting two units by 2 periods is $4c_v$ and the cost of expediting three units by 30 periods is $90c_v$. Although the variable expediting cost alone does not always allow for an accurate representation of actual expediting cost, we include it in our model to capture variable expediting costs if they are present and because essentially all other expediting models of the literature include it. Unlike previous research, we also include fixed expediting cost and expediting cost per batch and order in our model, because such costs are relevant in many real-world situations.

The fixed expediting cost c_f can be used to model situations where a fixed cost occurs each time one or multiple units are expedited. It includes, for instance, the time required to contact the supplier and to negotiate expediting, as well as the costs of rearranging and speeding up deliveries.

Our model does not contain hard constraints on expediting capacity. To incorporate capacity constraints in the model, such as the base capacity of a pallet or container, the expediting cost per batch c_b can be used. If up to q units are expedited, expediting cost c_b is charged, independently of the number of units expedited. If the number of units that are expedited exceeds q but does not exceed $2q$, expediting cost of $2c_b$ is charged, etc. By setting $q = 1$, we can model situations where expediting cost is proportional to the number of units expedited, but independent of the number of periods by which they are expedited.

Finally, the expediting cost per order c_o can be used to model situations where the expediting cost depends on the number of orders expedited. For instance, consider a supplier that has scheduled the production of two open orders in two different months. Expediting units from these two orders requires rescheduling two orders, which can be more costly than rescheduling a single order.

Our cost model is quite versatile and allows the specification of four expediting cost components. Not all expediting cost components are necessarily relevant, and it might suffice to consider only a subset. In the literature, the attention has been on variable expediting cost ($c_v > 0$, $c_f = 0$, $c_b = 0$, $c_o = 0$), whereas at the equipment manufacturer (see §7) an important expediting cost component is the fixed expediting cost ($c_v \geq 0$, $c_f > 0$, $c_b = 0$, $c_o = 0$).

Inventory is controlled by a periodic order-up-to level inventory policy with a constant expediting level K. Figure 1 shows the sequence of events in each period. Orders are placed at the very end of a period, whereas outstanding orders arrive at the beginning of a period. Therefore, the orders that are scheduled to arrive at the beginning of period t were placed $L + 1$ periods before. No events take place between the end of a period and the beginning of the next period and costs are the same if we assume that orders are placed at the very beginning of a period. However, we assume that orders are placed at the end of a period, because it later simplifies notation substantially (e.g., the readability of the cost function is significantly better if we can refer to the $(L + 1)$-dimensional order pipeline at the end of a period, instead of referring to the joint distribution of an L-dimensional order pipeline at the end of a period and the period's demand).

The sequence of events in each period is as follows. At the beginning of period t, we receive the order that was placed $L+1$ periods before, minus the number of units already expedited from that order. Simultaneously, we receive the units expedited in period $t - L_n$. Next, we determine the current expeditable pipeline stock, P_t^e, i.e., the total number of units that can potentially be expedited. If P_t^e exceeds the expediting level K, we expedite the $P_t^e - K$ oldest units, i.e., the $P_t^e - K$ units with the shortest residual lead times. Then, we observe and fill the demand D_t of the current period. We back-order excess demand and determine the on-hand inventory level, I_t; the back-order level, B_t; the total number of expedited lead time periods, C_t^v; the indicator for expedited units, C_t^f; the number of expedited batches, C_t^b; and the number of orders from which units are expedited, C_t^o; and we compute

Figure 1 **Sequence of Events**

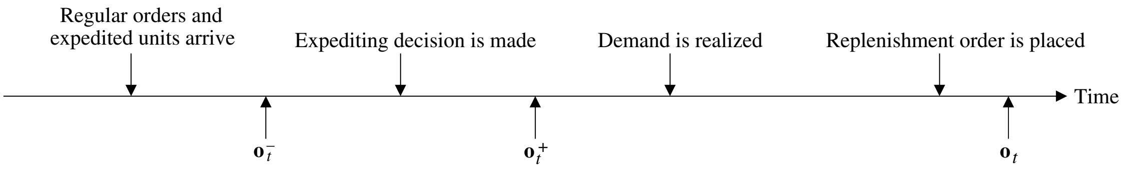
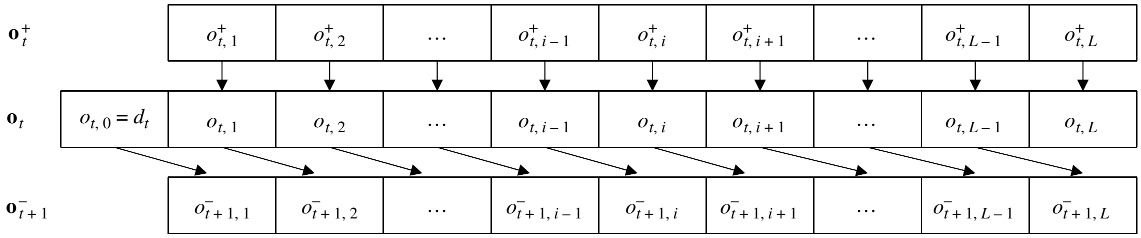

Regular orders and expedited units arrive
Expediting decision is made
Demand is realized
Replenishment order is placed
Time
$\mathbf{o}_t^-$
$\mathbf{o}_t^+$
$\mathbf{o}_t$

Figure 2 **Relationship Between the Outstanding Orders at the Relevant Points in Time**



the total cost of the period, $hI_t + bB_t + c_v C_t^v + c_f C_t^f + c_b C_t^b + c_o C_t^o$. Finally, we place a replenishment order to bring the inventory position (on-hand inventory minus back-orders plus outstanding orders) up to S. Because the order-up-to level is stationary, we reorder at the end of each period the demand of that period. The decision variables are the order-up-to level S and the expediting level K.

Note that we are using an inventory policy with a stationary expediting level, where expediting decisions are based on the total number of outstanding orders. This policy is not necessarily optimal because expediting levels that depend on the entire state-space of the inventory system, i.e., on the outstanding orders of individual periods and the back-order and inventory levels, can result in lower expected cost. However, as we will show in §6, the optimal policy can only be determined for small problems and is complex to implement, whereas the stationary expediting policy that we analyze captures a large portion of the potential savings, is easy to implement, and allows for solving problems of realistic sizeefficiently.

4. Inventory System Analysis

We first derive expressions for the pipeline stock distribution, which we then use to determine the objective function value. For our analyses, the outstanding orders at three points in time are relevant (Figure 1): after regular orders and expedited units have arrived ($\mathbf{o}_t^-$), after the expediting decision has been made ($\mathbf{o}_t^+$), and at the end of the period after the replenishment order has been made ($\mathbf{o}_t$). We next derive the corresponding distributions.

Let $o_{t,i}^+$ denote the number of units outstanding from the order that was placed in period $t-i$, after orders have been expedited in period t. The vector $\mathbf{o}_t^+ = (o_{t,1}^+, o_{t,2}^+, \ldots, o_{t,L}^+)$ represents the set of outstanding orders after the expediting decision, and $\sum_{i=1}^{l} o_{t,i}^+$ is the quantity outstanding from the previous l orders. Similarly, $\mathbf{o}_t = (o_{t,0}, o_{t,1}, o_{t,2}, \ldots, o_{t,L})$ denotes the set of outstanding orders at the end of period t, and $\mathbf{o}_t^- = (o_{t,1}^-, o_{t,2}^-, \ldots, o_{t,L}^-)$ denotes the set of outstanding orders before the expediting decision. Figure 2 illustrates the relationships between $\mathbf{o}_t^+$, $\mathbf{o}_t$, and $\mathbf{o}_{t+1}^-$. We denote the corresponding random variables by capital letters.

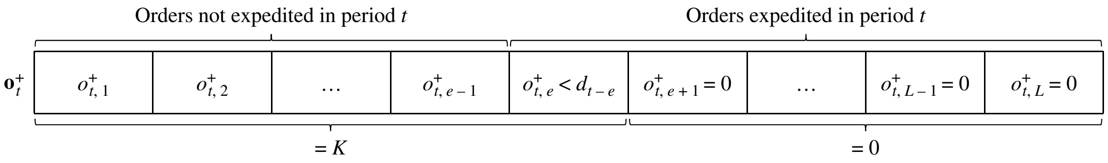
Determining the pipeline stock distribution with a positive nonexpeditable lead time, i.e., $L_n > 0$, is rather complex. Therefore, we first analyze the pipeline stock distribution for $L_n = 0$ and then build on the insights gained from the analysis to determine the pipeline stock distribution for $L_n > 0$.

4.1. Instantaneous Arrival of Expedited Units

If no outstanding unit has been expedited, then $o_{t,i}^+ = d_{t-i}$ for $i = 1, \ldots, L$. If some outstanding units have been expedited, then the oldest orders were expedited first and the outstanding orders have the structure shown in Figure 3, where $o_{t,e}^+$ denotes the youngest order expedited in period t. Orders $o_{t,1}^+, \ldots, o_{t,e-1}^+$ have not been expedited in period t and orders $o_{t,e+1}^+, \ldots, o_{t,L}^+$ have been expedited completely.

Now consider the l most recent outstanding orders $o_{t,1}^+, \ldots, o_{t,l}^+$. If the orders contain an expedited order, they must contain $o_{t,e}^+$. If the l most recent outstanding orders contain an order that was expedited in period t, the number of units outstanding from the most recent l orders is $\sum_{i=1}^{l} o_{t,i}^+ = \sum_{i=1}^{e} o_{t,i}^+ = K$. Proposition 1 provides information about the distribution of the outstanding units. All proofs are contained in the online supplement (available at http://dx.doi.org/10.1287/msom.2014.0506).

Figure 3 **Set of Outstanding Orders If Units Have Been Expedited in Period** *t*



Proposition 1. *If the number of units outstanding from the previous $l \in \{1, \ldots, L\}$ orders after the expediting decision is $\sum_{i=1}^{l} o^+_{t,i} < K$, then none of the previous l orders has been expedited, i.e., $o^+_{t,i} = d_{t-i}$ for $i = 1, \ldots, l$.*

Next, we determine the distribution of $\sum_{i=1}^{l} O^+_{t,i}$, i.e., the distribution of the number of units outstanding from the previous l orders after expediting. From Proposition 1, it follows that

$$P\left(\sum_{i=1}^{l} O^+_{t,i} = k\right) = P\left(\sum_{i=1}^{l} D_{t-i} = k\right) = f^l_D(k), \quad \text{for } k < K, \tag{2}$$

where $f^l_D(\cdot)$ denotes the l-fold convolution of the demand distribution.

The probability that the condition of Equation (2) holds, i.e., the probability that fewer than K units are outstanding from the previous l orders after expediting, is $P(\sum_{i=1}^{l} O^+_{t,i} < K) = F^l_D(K-1)$. If the condition does not hold, then exactly K units are outstanding after expediting. The corresponding probability is $P(\sum_{i=1}^{l} O^+_{t,i} = K) = 1 - F^l_D(K-1)$.

The steady-state distribution of the number of units outstanding from the previous l orders after expediting can be summarized as follows:

$$f^l_{O^+}(k) = P\left(\sum_{i=1}^{l} O^+_{t,i} = k\right) = \begin{cases} f^l_D(k) & \text{for } k < K, \\ 1 - F^l_D(K-1) & \text{for } k = K, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

After units have been expedited and before the end of the period, demand occurs and the demanded quantity is reordered. The distribution of the number of units outstanding from the previous l orders at the end of the period is the convolution of the distribution function of $\sum_{i=1}^{l} O^+_i$ and the demand of the period, i.e.,

$$f^l_O(k) = P\left(\sum_{i=0}^{l} O_{t,i} = k\right) = P\left(\sum_{i=1}^{l} O^+_{t,i} + D_t = k\right) = (f^l_{O^+} \star f_D)(k), \tag{4}$$

where the symbol "$\star$" indicates a convolution.

Before the expediting decision of period $t+1$ is made, the nonexpedited outstanding units from the order placed in period $t-L$ arrive. Therefore, the distribution of the number of units outstanding from the previous l orders before expediting corresponds to the distribution of the number of units outstanding from the previous $l-1$ orders at the end of the period, i.e.,

$$f^l_{O^-}(k) = P\left(\sum_{i=1}^{l} O^-_{t+1,i} = k\right) = P\left(\sum_{i=0}^{l-1} O_{t,i} = k\right) = f^{l-1}_O(k). \tag{5}$$

To compute the expected expediting cost, we next determine the distribution of $O^+_{t,l}$, given that $\sum_{i=1}^{l-1} O^+_{t,i} = m$, i.e., the distribution of the number of units outstanding after expediting from the order placed l periods before, given that the number of units outstanding from the previous $l-1$ orders is m. We start with the probability that fewer than $[K-m]^+$ units are outstanding after expediting from the order placed l periods before, i.e., $k < [K-m]^+$. From Proposition 1, it follows that

$$P\left(O^+_{t,l} = k \,\middle|\, \sum_{i=1}^{l-1} O^+_{t,i} = m\right) = P\left(D_{t-l} = k \,\middle|\, \sum_{i=1}^{l-1} D_{t-i} = m\right) = f_D(k), \quad \text{for } k < [K-m]^+. \tag{6}$$

The probability that the condition of Equation (6) holds is

$$P\left(O^+_{t,l} < [K-m]^+ \,\middle|\, \sum_{i=1}^{l-1} O^+_{t,i} = m\right) = F_D([K-m]^+ - 1).$$

If the condition does not hold, then exactly $[K-m]^+$ units are outstanding from the order placed l periods before. The corresponding probability is

$$P\left(O^+_{t,l} = [K-m]^+ \,\middle|\, \sum_{i=1}^{l-1} O^+_{t,i} = m\right) = 1 - F_D([K-m]^+ - 1).$$

Thus,

$$f_{O^+_l}\left(k \,\middle|\, \sum_{i=1}^{l-1} O^+_i = m\right) = P\left(O^+_{t,l} = k \,\middle|\, \sum_{i=1}^{l-1} O^+_{t,i} = m\right) = \begin{cases} f_D(k) & \text{for } k < [K-m]^+, \\ 1 - F_D(k-1) & \text{for } k = [K-m]^+, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Using Equations (3) and (7), we obtain

$$f^l_{O^+, O^+_l}(m, k) = P\left(\sum_{i=1}^{l} O^+_{t,i} = m \cap O^+_{t,l} = k\right) = f_{O^+_l}\left(k \,\middle|\, \sum_{i=1}^{l-1} O^+_i = m-k\right) f^{l-1}_{O^+}(m-k). \tag{8}$$

The joint distribution of the number of units outstanding at the end of the period from the previous l orders and the number of units outstanding from the order placed l periods before is the convolution of the joint distribution function of $\sum_{i=1}^{l} O_i^+$ and O_l^+, and the demand of the period, i.e.,

$$\begin{aligned} f^l_{O,O_l}(m,k) &= P\left(\sum_{i=0}^{l} O_{t,i} = m \cap O_{t,l} = k\right) \\ &= \sum_{n=0}^{m-k} f^l_{O^+,O_l^+}(m-n,k) f_D(n). \qquad (9) \end{aligned}$$

The joint distribution of the number of units outstanding before expediting from the previous l orders and the number of units outstanding from the order placed l periods before corresponds to the joint distribution of the number of units outstanding at the end of the period from the previous $l-1$ orders and the number of units outstanding from the order placed $l-1$ periods before, i.e.,

$$\begin{aligned} f^l_{O^-,O_l^-}(m,k) &= P\left(\sum_{i=1}^{l} O^-_{t+1,i} = m \cap O^-_{t+1,l} = k\right) \\ &= f^{l-1}_{O,O_{l-1}}(m,k). \qquad (10) \end{aligned}$$

4.2. Noninstantaneous Arrival of Expedited Units

If expedited units do not arrive instantaneously, but after a nonexpeditable lead time of $L_n < L$ periods, then expediting is only feasible for orders placed within the previous $L_e = L - L_n$ periods.

We can partition the outstanding orders into two sets (Figure 4). In the figure, expediting options are indicated by arrows. To derive the distribution of the number of units outstanding after expediting in period t, we consider the distribution of the units outstanding in period $t - L_n$. Then, we can neglect all units in the nonexpeditable set $o^+_{t-L_n,L_e+1}, \ldots, o^+_{t-L_n,L_e+L_n}$, because these units are neither outstanding in period t, nor do they have an impact on any unit outstanding in period t. The distribution of the units outstanding in the expeditable set $o^+_{t-L_n,1}, \ldots, o^+_{t-L_n,L_e}$ corresponds to the distribution of the units outstanding in the model with instantaneous arrival of expedited units and regular lead time $L = L_e$, because the expediting decision depends only on the distribution of the units outstanding from the previous L_e orders. Furthermore, we can neglect order expediting between period $t - L_n$ and period t, because every unit expedited after period $t - L_n$ will still be outstanding in period t.

Therefore, the steady state distribution of the number of units outstanding after expediting with expeditable lead time L_e and nonexpeditable lead time L_n is the convolution of the distribution function of the number of units outstanding after expediting in the model with instantaneous arrival of expedited units and regular lead time $L = L_e$ and the L_n-fold convolution of the demand distribution, i.e.,

$$f^{L_e,L_n}_{O^+}(k) = (f^{L_e}_{O^+} \star f^{L_n}_D)(k).$$

The distribution of the number of units outstanding at the end of the period,

$$f^{L_e,L_n}_{O}(k) = (f^{L_e,L_n}_{O^+} \star f_D)(k) = (f^{L_e}_{O^+} \star f^{L_n+1}_D)(k), \qquad (11)$$

and the distribution of the number of units outstanding before expediting,

$$f^{L_e,L_n}_{O^-}(k) = \begin{cases} f^{L_e-1,0}_{O}(k) & \text{for } L_n = 0, \\ f^{L_e,L_n-1}_{O}(k) & \text{for } L_n > 0, \end{cases}$$

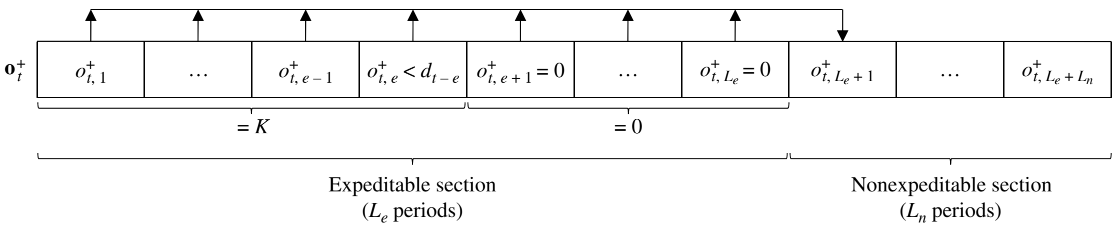
can be obtained along the same line of arguments as in the instantaneous arrival version of the model.

Note that the distribution of the number of units outstanding from the previous L_e orders is independent of the nonexpeditable lead time L_n, i.e., $f^{l,0}_{O^+}(k) = f^l_{O^+}(k)$ holds for $l \le L_e$. The joint distributions of the number of units outstanding from the previous l orders and the number of units outstanding from the order placed l periods before (Equations (8)–(10)) are only required to compute the expected expediting cost. Therefore, for $l \le L_e$, Equations (8)–(10) also hold for positive L_n.

4.3. Objective Function

Our objective is to minimize the sum of expected inventory holding, back-order penalty, and expediting cost (Equation (1)). *Expected inventory holding* and *back-order penalty cost* can be computed based on the distribution of the outstanding orders at the end of

Figure 4 **Structure of Outstanding Orders If Expedited Units Arrive After a Nonexpeditable Lead Time**



the period (Equation (11)) as $h\sum_{k=0}^{S}(S-k)f_O^{L_e,L_n}(k)$ and $b\sum_{k=S}^{\infty}(k-S)f_O^{L_e,L_n}(k)$, respectively. We next determine expressions for the four expected expediting cost components.

Variable expediting cost is incurred for each unit and period of expediting. Consider the order that was placed in period $t-l$ and is scheduled to arrive in period $t-l+L_e+L_n+1$. If we expedite units of this order, then they arrive in period $t+L_n$, i.e., L_e-l+1 periods before the regularly scheduled arrival period. Let e_l denote the number of units of this order that are expedited in period t. Then, the variable expediting cost is $c_v(L_e-l+1)e_l$ and the expected variable expediting cost in a period is $E[\sum_{l=1}^{L_e}c_v(L_e-l+1)e_l]$.

It is difficult to compute expected values of the individual e_l, but it is relatively easy to compute $E[\sum_{i=1}^{l}e_i]$ (see below). Therefore, we reformulate the expression for the expected variable expediting cost:

$$E\left[\sum_{l=1}^{L_e}c_v(L_e-l+1)e_l\right]$$
$$=E[c_v(L_e e_1+(L_e-1)e_2+\cdots+1e_{L_e})]=c_v\sum_{l=1}^{L_e}E\left[\sum_{i=1}^{l}e_i\right].$$

The term $E[\sum_{i=1}^{l}e_i]$ represents the expected number of units from the previous l orders that are expedited in a period. We can compute this quantity based on the number of units outstanding from the previous l orders before expediting. If $k>K$ units are outstanding before expediting, then $k-K$ units are expedited. The probability that k units from the previous $l\le L_e$ orders are outstanding before expediting is $f_{O^-}^{l}(k)$ (Equation (5)). Therefore, the expected number of units from the previous l orders that are expedited in a period is $E[\sum_{i=1}^{l}e_i]=\sum_{k=K}^{\infty}(k-K)f_{O^-}^{l}(k)$, and the expected variable expediting cost is

$$c_v\sum_{l=1}^{L_e}\sum_{k=K}^{\infty}(k-K)f_{O^-}^{l}(k).$$

Fixed expediting cost is incurred for each period in which orders are expedited. The probability that expediting occurs in a period corresponds to the probability that more than K units are outstanding from the previous L_e orders before expediting, i.e., $\sum_{k=K+1}^{\infty}f_{O^-}^{L_e}(k)$. Therefore, the expected fixed expediting cost is $c_f\sum_{k=K+1}^{\infty}f_{O^-}^{L_e}(k)=c_f[1-F_{O^-}^{L_e}(K)]$.

Similarly, we can determine the *expediting cost per batch*, which is incurred for each batch of up to q units that are expedited. The probability that at least one batch is required corresponds to the probability that more than K units are outstanding from the previous L_e orders before expediting, i.e., $\sum_{k=K+1}^{\infty}f_{O^-}^{L_e}(k)$. The probability that at least two batches are required corresponds to the probability that more than $K+q$ units are outstanding from the previous L_e orders before expediting, i.e., $\sum_{k=K+q+1}^{\infty}f_{O^-}^{L_e}(k)$. More generally, the probability that at least i batches are required corresponds to the probability that more than $K+(i-1)q$ units are outstanding from the previous L_e orders before expediting, i.e., $\sum_{k=K+(i-1)q+1}^{\infty}f_{O^-}^{L_e}(k)$. Therefore, the expected expediting cost per batch is $c_b\sum_{i=0}^{\infty}[1-F_{O^-}^{L_e}(K+i\cdot q)]$.

Finally, we determine the *expediting cost per order* from which units are expedited. The variable e_l denotes the number of units that are expedited in period t from the order that was placed in period $t-l$. The probability that units of this order are expedited in period t is $P(e_l>0)=[1-P(e_l=0)]$, and the expected expediting cost per order can be written as $c_o\sum_{l=1}^{L_e}[1-P(e_l=0)]$.

The term $P(e_l=0)$ corresponds to the probability that no units of the order that was placed in period $t-l$ are expedited in period t. Units from the order that was placed l periods before are expedited only if $m>K$ units are outstanding from the previous $l\le L_e$ orders and $k>0$ units are outstanding from the order that was placed l periods before. The probability that m units are outstanding from the previous $l\le L_e$ orders before expediting and k units are outstanding from the order that was placed l periods before is $f_{O^-,O_l^-}^{l}(m,k)$ (Equation (10)). Therefore, the probability that no units of the order that was placed in period $t-l$ are expedited in period t is

$$P(e_l=0)=\sum_{m=0}^{\infty}\sum_{k=0}^{m}\mathbb{1}_{(m\le K\cup k=0)}f_{O^-,O_l^-}^{l}(m,k)$$
$$=\sum_{m=0}^{K}\sum_{k=1}^{m}f_{O^-,O_l^-}^{l}(m,k)+\sum_{m=0}^{\infty}f_{O^-,O_l^-}^{l}(m,0).$$

The left term of the last expression is the probability that units are outstanding from the order that was placed l periods before, but the total number of units that are outstanding from the previous $l\le L_e$ orders is below K. The right term of the last expression is the probability that no units are outstanding from the order that was placed l periods before. Thus, the expected expediting cost per order is

$$c_o\left[L_e-\sum_{l=1}^{L_e}\left[\sum_{m=0}^{K}\sum_{k=1}^{m}f_{O^-,O_l^-}^{l}(m,k)+\sum_{m=0}^{\infty}f_{O^-,O_l^-}^{l}(m,0)\right]\right].$$

Our objective function can be written as

$$Z(S,K)=h\sum_{k=0}^{S}(S-k)f_O^{L_e,L_n}(k)+b\sum_{k=S}^{\infty}(k-S)f_O^{L_e,L_n}(k)$$
$$+c_v\sum_{l=1}^{L_e}\sum_{k=K}^{\infty}(k-K)f_{O^-}^{l}(k)+c_f[1-F_{O^-}^{L_e}(K)]$$
$$+c_b\sum_{i=0}^{\infty}[1-F_{O^-}^{L_e}(K+i\cdot q)]$$

$$+ c_o\Bigg[L_e - \sum_{l=1}^{L_e}\Bigg[\sum_{m=0}^{K}\sum_{k=1}^{m} f^{l}_{O^-,\,O^-_l}(m,k) + \sum_{m=0}^{\infty} f^{l}_{O^-,\,O^-_l}(m,0)\Bigg]\Bigg]. \quad (12)$$

5. Solution Approach

We first develop a solution of a simplified version of the model that considers only variable expediting cost c_v and ignores the expediting cost components c_f, c_b, and c_o. In §5.1, we show how the optimal solution of the simplified model can be determined very efficiently by exploiting a convexity property. In §5.2, we use these findings as a building block in the solution approach of the full model.

5.1. Simplified Model

The simplified objective function that we analyze in this section considers variable expediting cost only:

$$Z^s(S,K) = h\sum_{k=0}^{S}(S-k)f_O^{L_e,\,L_n}(k) + b\sum_{k=S}^{\infty}(k-S)f_O^{L_e,\,L_n}(k) + c_v\sum_{l=1}^{L_e}\sum_{k=K}^{\infty}(k-K)f^{l}_{O^-}(k).$$

The first- and second-order finite differences of the function $Z^s(S,K)$ with respect to S are defined as $\Delta_S Z^s(S,K) = Z^s(S+1,K) - Z^s(S,K)$ and $\Delta_S^2 Z^s(S,K) = \Delta_S Z^s(S,K) - \Delta_S Z^s(S-1,K)$, respectively. The finite differences $\Delta_K Z^s(S,K)$ and $\Delta_K^2 Z^s(S,K)$ are defined correspondingly. The first- and second-order finite differences of $Z^s(S,K)$ with respect to S and K are defined as $\Delta_{S,\,K} Z^s(S,K) = Z^s(S+1,K+1) - Z^s(S,K)$ and $\Delta^2_{S,\,K} Z^s(S,K) = \Delta_{S,\,K} Z^s(S,K) - \Delta_{S,\,K} Z^s(S-1,K-1)$.

Next, we derive structural results of the objective function $Z^s(S,K)$ that we utilize to design an algorithm for computing the optimal order-up-to level S^* and the optimal expediting level K^*. First, we analyze $Z^s(S,K)$ with respect to changes in S for a fixed K and derive a lower and an upper bound on S^*. Then, we show that there exists a linear relationship between S^* and K^*, which allows us to limit the search space to one dimension. We also prove convexity on the linear search space, which allows us to use efficient optimization approaches for finding the optimal solution. Finally, we describe the optimization algorithm for the simplified model.

5.1.1. Bounds on the Optimal Order-Up-To Level. We start our analysis by examining the effect of the order-up-to level S on expected cost $Z^s(S,K)$ for a given expediting level K. The first-order finite difference with respect to S is $\Delta_S Z^s(S,K) = (b+h)F_O^{L_e,\,L_n}(S) - b$, and the second-order finite difference is $\Delta_S^2 Z^s(S,K) = (b+h)f_O^{L_e,\,L_n}(S)$. The derivations of these and all following finite difference functions are contained in the online supplement.

Because $\Delta_S^2 Z^s(S,K) \geq 0$, the expected cost function is convex in S, and we can set the first-order finite difference equal to zero and solve for S to obtain the optimal order-up-to level

$$S^*(K) = (F_O^{L_e,\,L_n})^{-1}\frac{b}{b+h}, \quad (13)$$

where $F_O^{L_e,\,L_n}(\cdot)$ is defined by Equation (11) and depends on K. Equation (13) allows us to determine the optimal order-up-to level $S^*(K)$ for a given expediting level K. To obtain a lower and an upper bound on S^*, we rely on Proposition 2. We note that we use the terms "increasing" and "decreasing" to refer to nondecreasing and nonincreasing functions. If a function is strictly increasing or decreasing, then we state this explicitly. Similarly, we use the terms "positive" and "negative" to refer to nonnegative and nonpositive values.

Proposition 2. *The optimal order-up-to level $S^*(K)$ is increasing in K.*

Because $S^*(K)$ is increasing in K, a lower bound on the optimal order-up-to level is $S_{\min} = S^*(0)$. For $K = 0$, each new order is expedited directly into the nonexpeditable section of the order pipeline (see Figure 4). Therefore, the only outstanding orders at the end of a period are those in the nonexpeditable section of the order pipeline and those that were placed in that period. Because we order in each period the demand of that period, we can compute the lower bound $S_{\min}$ based on the inverse of the distribution function of the demand over $L_n + 1$ periods:

$$S_{\min} = (F_D^{L_n+1})^{-1}\frac{b}{b+h}. \quad (14)$$

To determine an upper bound on the optimal order-up-to level, we analyze the optimal order-up-to level for $K \to \infty$, i.e., $S_{\max} = \lim_{K\to\infty} S^*(K)$. Then, no order is expedited and we execute a standard periodic review inventory policy. The upper bound on the order-up-to level can be computed based on the inverse of the distribution function of the demand over $L_e + L_n + 1$ periods:

$$S_{\max} = (F_D^{L_e+L_n+1})^{-1}\frac{b}{b+h}. \quad (15)$$

5.1.2. Quasiconvexity of the Expediting Level. Consider the first-order finite difference of the expected cost function with respect to the expediting level K,

$$\Delta_K Z^s(S,K) = [1 - F_D^{L_e}(K)]\big[b - c_v - (b+h)F_D^{L_n+1}(S-K-1)\big]. \quad (16)$$

Figure 5 Effect of K on Z^s, OH, BO, and EX for Given S

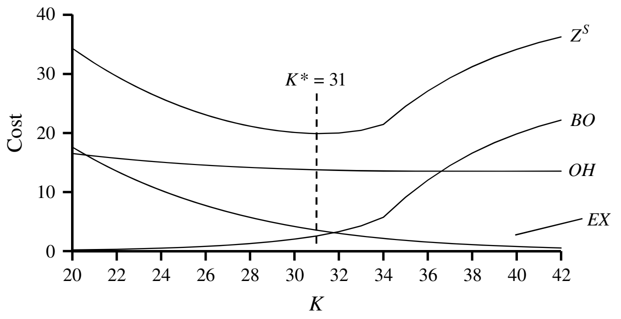


Figure 5 shows how expected cost (Z^s), expected on-hand inventory holding cost (OH), expected back-order penalty cost (BO), and expected expediting cost (EX) depend on the expediting level K for an example with order-up-to level $S = 34$, unit inventory holding cost $h = 1$, unit back-order penalty cost $b = 50$, variable expediting cost $c_v = 5$, lead time $L = L_e = 20$, and negative binomial distributed demand with mean $\mu = 1$ and standard deviation $\sigma = 2$. The figure shows that the expected cost function is not convex in K. However, we can show that it is quasiconvex in K and that there exists a linear relationship between the optimal order-up-to level S^* and the optimal expediting level K^* (Proposition 3).

Proposition 3. *For $b > c_v$ and a given order-up-to level S, $Z^s(S, K)$ is quasiconvex in K and the optimal expediting level is*

$$K^*(S) = S - \left(F_D^{L_n+1}\right)^{-1} \frac{b - c_v}{b + h}. \qquad (17)$$

Proposition 3 considers the case where $b > c_v$, i.e., where the unit back-order penalty cost is greater than the variable expediting cost. If $b \le c_v$, then $\Delta_K Z^s(S, K)$ is negative and the optimal expediting level is $K^* = \infty$; i.e., we never expedite and the optimal order-up-to-level is $S^* = S_{\max}$ (Equation (15)).

Equation (17) can be nicely interpreted. It states that there exists a fixed difference of $(F_D^{L_n+1})^{-1}((b - c_v)/(b + h))$ between the order-up-to level and the optimal expediting level. In an inventory system, this value represents the quantity to which we increase the inventory level each period when we make the expediting decision. In other words, when we decide on the number of units to expedite, we expedite a quantity such that all back orders are filled and the inventory level is raised to $(F_D^{L_n+1})^{-1}((b - c_v)/(b + h))$.

To find the optimal solution (S^*, K^*), we could evaluate $Z^s(S, K^*(S))$ for all $S = S_{\min}, \ldots, S_{\max}$. However, we can improve the efficiency of the solution approach by utilizing the convexity property stated in Proposition 4.

Proposition 4. *The expected cost $Z^s(S, K^*(S))$ is convex in S for $S_{\min} \le S \le S_{\max}$.*

Because $Z^s(S, K^*(S))$ is convex in S, we can use standard convex optimization approaches for finding the optimal solution more efficiently. The first finite difference $\Delta_{S,K} Z^s(S, K^*(S))$ can be calculated much faster than $Z^s(S, K^*(S))$, because the evaluation of the expected variable expediting cost is computationally expensive. Therefore, we use a root-finding algorithm that relies on $\Delta_{S,K} Z^s(S, K^*(S))$ as opposed to $Z^s(S, K^*(S))$ to find the optimal solution.

5.2. Full Model

Our optimization approach for the full model builds on the optimal solution of the simplified model. From Equation (12), we can see that the expected expediting cost components depend on the expediting level K, but not on the order-up-to level S. This is intuitive, because the expediting decision depends on the distribution of the open orders in the order pipeline only. The order pipeline distribution is independent of the order-up-to level S, because the demand of the period is ordered at the end of a period. The optimal order-up-to level $S^*(K)$ for a given expediting level K is independent of c_v, c_f, c_b, and c_o, and $S^*(K)$ corresponds to the optimal order-up-to level for a given expediting level of the simplified model, i.e., $S^*(K) = (F_O^{L_e, L_n})^{-1}(b/(b + h))$. Using $S^*(K)$, we can limit the search space of the objective function $Z(S, K)$ to one dimension, the expediting level K.

Next, we derive bounds on the optimal expediting level. The parameter K^* denotes the optimal expediting level of the full model with expediting costs c_v, c_f, c_b, and c_o. Let $\tilde{K}^*$ denote the optimal expediting level if cost $c_v + c_f + c_b + c_o$ is charged for each unit and period of expediting with all other expediting costs set to zero. $Z(S, K)$ and $\tilde{Z}(S, K)$ denote the corresponding objective functions.

Inequality $Z(S^*(K^*), K^*) \le Z(S^*(\tilde{K}^*), \tilde{K}^*)$ holds, because the optimal order-up-to level for a given expediting level K is $S^*(K)$, and $Z(S^*(K), K)$ is minimized for $K = K^*$. With a similar argument it can be seen that $\tilde{Z}(S^*(\tilde{K}^*), \tilde{K}^*) \le \tilde{Z}(S^*(K^*), K^*)$ holds, which implies that $\tilde{Z}(S^*(K^*), K^*) - Z(S^*(K^*), K^*) \ge \tilde{Z}(S^*(\tilde{K}^*), \tilde{K}^*) - Z(S^*(\tilde{K}^*), \tilde{K}^*)$. Because $\tilde{Z}(S^*(K), K) - Z(S^*(K), K)$ is decreasing in K (see proof of Proposition 5 for details), we obtain the following proposition:

Proposition 5. *If $\tilde{K}^* < \infty$, then*

$$0 \le K^* \le \tilde{K}^*.$$

Proposition 5 states that the optimal expediting level is higher if the expediting cost components c_f, c_b, or c_o are charged for each unit and period of expediting (in addition to c_v), instead of being charged for the quantities to which they are actually related. Because only variable expediting cost is charged, $\tilde{K}^*$ can be calculated by the optimization algorithm for the simplified model and serves as an upper bound on K^*. Note that in situations in which $\tilde{K}^*$ is not finite, we do not obtain a bound on $\tilde{K}^*$ and must search for K^* up to the maximum demand over the lead time.

To solve the full model, we first determine an upper bound for K^* using Proposition 5 or the maximum demand over the lead time. Then, we can evaluate $Z(S^*(K), K)$ with expediting costs c_v, c_f, c_b, and c_o for all $K = 0, \ldots, \tilde{K}^*$ to find the optimal solution of the full model, (S^*, K^*).

We have shown that the full model can be solved similarly to the simplified model. To solve the full model, we can rely on the results that we derived for the simplified model and reduce the search for the optimal solution to a one-dimensional optimization of a convex function. As we will show next, this allows us to solve even large models very efficiently, and considerable cost savings can be achieved.

6. Numerical Evaluation of Performance of Expediting Policy

We analyze the performance of the *expediting policy* (EP) by comparing it with the performances of the standard order-up-to level policy and two optimal policies. The *standard policy* (SP) uses an order-up-to level of $S^{\mathrm{SP}} = (F_D^{L+1})^{-1}(b/(b+h))$ and does not expedite any order. The optimal policies make expediting and replenishment decisions based on the inventory level and the open orders in each lead time period. We consider two order-expediting options for the optimal policies. Under the *optimal first-come, first-served expediting policy* (OP-FCFS), open orders are expedited in the sequence in which they were placed. Under the *optimal free expediting policy* (OP-FREE), any open order can be expedited. To compute the optimal policies, we formulate the problem as a dynamic program and solve it using single-chain value iteration (White 1963, Odoni 1969, Bertsekas 2007).

We define a base case that uses the data of an average part of the equipment manufacturer. The average part has a lead time of 55 days, and the company uses a review period of one day. The standard policy and the expediting policy can solve problems with such lead times, but the dynamic program is computationally tractable only for single-digit lead times. We therefore use in this section a review period of 11 days, such that the lead time of an average part is five review periods and the optimal solution can be computed. The average part has a demand rate of $\lambda = 1.21$ units per period, unit back-order penalty cost of $b = 550$ per period, unit inventory holding cost of $h = 11$ per period, and fixed expediting cost of $c_f = 45$. In the base case, we set the other expediting costs equal to zero ($c_v = c_b = c_o = 0$) and analyze them in a separate analysis. For the numerical experiments, we use Poisson demand and a fraction for the nonexpeditable lead time of 20% of the regular lead time. Then, for example, a part with a regular lead time of 10 review periods arrives two review periods after the period in which it has been expedited. The results for the base case and for cases in which we varied each of the problem parameters are shown in Table 1. The algorithms were implemented in C++, and all experiments were conducted on a PC running at 2.4 GHz with 8 GB of RAM. For problems for which $\tilde{K}^*$ is infinite, we use $(F_D^{L_e+L_n+1})^{-1}(1-\epsilon)$ with $\epsilon = 10^{-6}$ as an upper bound.

6.1. Expediting Policy vs. Standard Policy

The numerical results indicate that substantial savings can be achieved by expediting open orders (Table 1). The expected costs under the expediting policy are, on average, 22.4% lower than under the standard policy. The numerical results also indicate that the savings are particularly high if unit back-order penalty cost b, unit inventory holding cost h, demand rate λ, or lead time L are high or if the fixed expediting cost c_f is low. Note that the percentage cost savings are not always increasing in the regular lead time L. This is because 20% of the regular lead time is nonexpeditable and lead times are integers. Therefore, the nonexpeditable lead time is $L_n = 0$ if the regular lead time is one to four, $L_n = 1$ if the regular lead time is five to nine, etc.... Each time the nonexpeditable lead time increases, the savings decrease.

Computing the solution of the standard policy requires negligible CPU time. Computing the solution of the expediting policy is also fast. The solution can be computed with less than one millisecond of CPU time for small problems, and even the largest problem with a lead time of 100 periods can be solved in less than one second.

6.2. Expediting Policy vs. Optimal FCFS Expediting Policy

Under the expediting policy, we only consider the number of open orders to decide if open orders should be expedited. This structure is not optimal, and the optimal FCFS expediting policy relies on the number of open orders in each lead time period.

Not all test cases could be solved by dynamic programming. For the cases that could be solved, the average cost of the optimal FCFS policy is 21.2% below the expected cost of the standard policy. For these cases, the average cost savings of the expediting policy compared to the expected cost of the standard policy are 16.5%, which indicates that the expediting policy can realize on average much of the savings potential that order expediting offers.

Table 1 **Performances of Standard Policy, Expediting Policy, and Optimal Policies**

	Parameter values					SP		EP					OP-FCFS			OP-FREE		
	λ	L	b	h	c_f	S^*	Z	S^*	K^*	Z	% sav.	Time (ms)	Z	% sav.	Time (ms)	Z	% sav.	Time (ms)
Base	1.21	5	550	11	45	13	79.98	11	6	67.33	15.8	1	62.85	21.4	21,840	60.86	23.9	851,273
λ	0.12					3	29.23	2	1	23.38	20.0	<1	22.70	22.4	1,975	22.70	22.4	7,470
	0.60					8	58.31	7	4	49.30	15.4	<1	45.90	21.3	7,954	45.35	22.2	146,625
	1.21					13	79.98	11	6	67.33	15.8	1	62.85	21.4	21,690	60.86	23.9	862,129
	6.03					49	169.18	43	24	133.83	20.9	16						
	12.05					90	236.09	80	46	177.88	24.7	61						
L		1				6	48.70	6	3	46.12	5.3	<1	45.43	6.7	154	45.43	6.7	180
		2				8	58.31	7	4	50.11	14.1	<1	48.49	16.8	167	47.80	18.0	297
		3				10	66.53	8	5	53.31	19.9	<1	50.27	24.4	1,111	48.75	26.7	12,091
		4				12	74.21	9	6	55.82	24.8	<1	51.62	30.4	9,843	49.17	33.7	474,957
		5				13	79.98	11	6	67.33	15.8	1	62.85	21.4	21,525	60.86	23.9	882,549
		6				15	85.62	13	8	69.45	18.9	2	63.73	25.6	270,233	61.15	28.6	44,054,316
		7				17	91.77	14	9	70.97	22.7	2						
		8				18	96.13	15	10	72.31	24.8	2						
		9				20	101.19	16	11	73.52	27.3	3						
		10				21	105.66	18	11	81.84	22.5	3						
		20				36	142.91	29	18	100.54	29.6	10						
		30				50	172.31	40	25	114.99	33.3	21						
		40				64	196.74	51	33	125.93	36.0	42						
		50				78	218.15	62	41	135.81	37.7	52						
		60				92	237.87	73	49	144.75	39.1	80						
		70				105	255.90	84	57	152.92	40.2	105						
		80				119	272.98	95	65	160.46	41.2	141						
		90				132	288.59	106	73	167.48	42.0	192						
		100				145	303.54	117	81	174.11	42.6	221						
b			55			10	46.52	9	7	43.90	5.6	1	39.37	15.4	29,048	38.11	18.1	986,418
			275			12	70.50	11	7	60.93	13.6	1	56.01	20.6	23,963	54.61	22.5	852,160
			550			13	79.98	11	6	67.33	15.8	1	62.85	21.4	21,543	60.86	23.9	852,980
			2,750			15	100.51	13	6	81.87	18.5	1	77.25	23.1	18,958	75.74	24.6	639,204
			5,500			16	108.41	13	6	87.03	19.7	1	82.88	23.5	17,696	80.51	25.7	727,403
h				1.1		16	10.84	15	10	10.30	5.0	1	9.73	10.3	18,959	9.68	10.7	661,423
				5.5		14	44.46	13	8	39.45	11.3	1	36.63	17.6	20,265	35.93	19.2	687,171
				11		13	79.98	11	6	67.33	15.8	1	62.85	21.4	21,357	60.86	23.9	859,265
				55		11	284.47	8	4	204.10	28.3	<1	195.16	31.4	28,958	193.57	32.0	984,930
				110		10	465.16	5	1	305.27	34.4	<1	305.10	34.4	31,459	303.96	34.7	14,418,066
c_f					4.5	13	79.98	7	1	51.85	35.2	<1	51.85	35.2	26,554	51.75	35.3	7,285,128
					22.5	13	79.98	10	4	61.70	22.9	1	58.83	26.4	24,104	57.42	28.2	894,158
					45	13	79.98	11	6	67.33	15.8	1	62.85	21.4	21,492	60.86	23.9	852,729
					225	13	79.98	13	10	77.13	3.6	1	71.89	10.1	20,182	70.97	11.3	681,013
					450	13	79.98	13	11	79.10	1.1	1	75.45	5.7	16,496	75.04	6.2	648,609
Average (only cases with optimal solution)											16.5	<1		21.2	27,901		22.8	3,172,902
Average (all cases)											22.4	24						

6.3. Optimal FCFS Expediting Policy vs. Optimal Free Expediting Policy

If orders do not have to be expedited FCFS, but any open order can be expedited, additional optimization potential exists. Compared to the standard policy, the optimal free expediting policy achieves average savings of 22.8%, i.e., 1.6% higher savings than the optimal FCFS expediting policy. The CPU times required to compute the optimal free expediting solutions are higher than for the optimal FCFS expediting solutions, because there are many more expediting options to analyze.

6.4. Effect of Expediting Cost Types on Cost Savings

In Table 1, we have considered only the fixed expediting cost c_f, and the upper row of Table 2 shows the aggregate results for the cases that could be solved optimally. The remaining rows show the cost savings for other types of expediting costs. From the first three rows of the table it can be seen that the cost savings are similar for fixed expediting cost ($c_f = 45$), expediting cost per batch ($c_b = 45$, $q = 3$ in the experiment), and expediting cost per order ($c_o = 45$). If

Table 2 **Average Cost Savings in Different Expediting Cost Settings**

Parameter values				Average cost savings (%)			Average CPU time (ms)		
c_v	c_f	c_b	c_o	EP	OP-FCFS	OP-FREE	EP	OP-FCFS	OP-FREE
0	45	0	0	16.5	21.2	22.8	<1	27,901	3,172,902
0	0	45	0	16.2	20.2	22.4	<1	26,780	2,532,168
0	0	0	45	15.4	18.7	22.7	1	24,479	3,282,911
55	45	0	0	6.3	6.7	6.8	1	21,418	1,572,347
55	0	45	0	6.3	6.6	6.7	1	21,291	1,571,389
55	0	0	45	6.1	6.4	6.7	1	21,409	1,580,830

Table 3 Performance Indicators for Different Demand Rates and Lead Times

	Base case	Higher demand	Shorter lead time	Higher demand and shorter lead time
Avg. $\mu =$	2.1	21.1	2.1	21.1
Avg. $L =$	43.3	43.3	21.6	21.6
Fraction of SKUs using expediting policy (%)	53.7	47.9	47.6	42.8
Overall fraction of demand expedited (%)	23.4	16.2	28.1	13.8
Average number of units expedited (per expediting)	1.17	1.47	1.16	1.43
Lead time reduction (per expedited unit) (%)	46.9	18.2	59.6	24.6
Cost savings (%)	34.6	32.8	34.2	30.9

variable expediting cost ($c_v = 55$) is included, the cost savings decrease, because it becomes more expensive to expedite.

The numerical experiments show that one can achieve considerable cost savings, but that the savings depend on the problem parameters. Companies handle a variety of parts, and to obtain a better understanding about the cost savings that are reasonable to expect in practice, we next apply the expediting policy to a portfolio of parts of the equipment manufacturer.

7. Application

We use data from the service division of the equipment manufacturer to quantify the potential cost savings of order expediting for a specific example and conduct sensitivity analyses to analyze how the results can be expected to generalize. We use data from 600 stock keeping units (SKUs) with an economic order quantity of one. The annual inventory holding cost rate is 15%, the unit back-order penalty cost is 50 times the unit inventory holding cost, and the nonexpeditable section of the lead time is 20%. We set the fixed expediting cost equal to the order setup cost of the company, i.e., $c_f = 45$, and use variable expediting cost of $c_v = 5$. The expediting cost per batch c_b and the expediting cost per order c_o can be neglected at the company, and we analyze them in a separate sensitivity analysis later. If the variance-to-mean ratio of the demand is one or below, we use a Poisson demand model; if the variance-to-mean ratio is above one, we use a negative binomial demand model.

The column "Base case" in Table 3 shows the results of the optimization. It indicates that 53.7% of the SKUs use an expediting policy. The fraction of demand that is expedited is 23.4%; i.e., approximately every fourth unit of demand is expedited. If open orders are expedited, an average of 1.17 units are expedited. Overall, order expediting reduces expected cost by 34.6%.

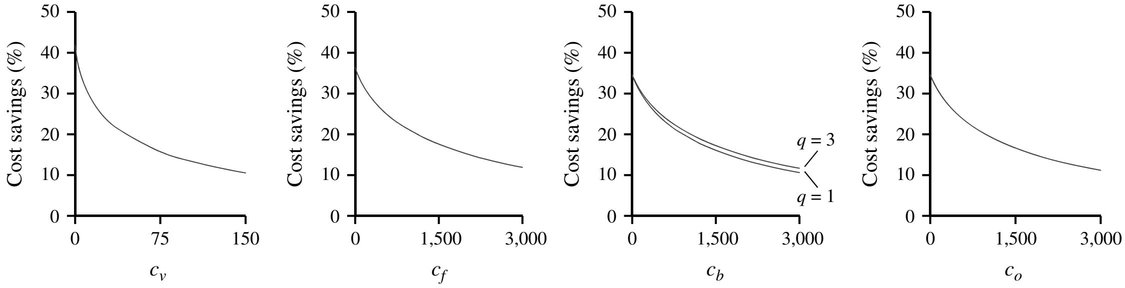
In the example, the average demand rate is 2.1 units per year and the average lead time is 43.3 days. In other industries, the average demand rate or the average lead time might be different, and Table 3 shows the results for situations with higher demand rates or shorter lead times.

We can observe that, in settings with high demand rates, fewer SKUs are expedited than in settings with low demand rates (second versus first and fourth versus third columns). The higher the demand rate, the better is safety stock utilized and the more beneficial it becomes to invest in safety stock compared to using order expediting. We can also observe that, in settings with short lead times, fewer SKUs are expedited than in settings with long lead times (third versus first and fourth versus second columns). The shorter the lead time, the smaller is the demand uncertainty over the lead time and the less likely it is that units must be expedited. Cost savings are similar across settings.

Figure 6 shows how expediting costs affect the cost savings. The graphs show that each expediting cost strongly affects the cost savings, but that, even for large values of the expediting cost, substantial savings can be achieved.

The third graph in Figure 6 also shows the effect of the batch size assumption on expected cost. Note that the cost savings for batch sizes of $q = 1$ to $q = 3$ are

Figure 6 Effect of Expediting Model Parameters on Cost Savings



similar, because the average number of units that are expedited per expediting is relatively small. Averaged over all parts and c_b values, we expedite 1.15 units for $q=1$ and 1.18 units for $q=3$. However, for parts with relatively high demand rates and high inventory holding and back-order penalty cost, the difference can be substantial. For instance, for a part with $\lambda=12.1$, $L=5$, $b=5{,}500$, $h=110$, and $c_b=45$, we expedite on average 6.52 units for $q=1$ and 7.96 units for $q=3$, with cost savings of 28.6% and 34.4%, respectively. In such situations, the effect of the batch size on expected cost is substantial, and choosing an appropriate batch size is important for such parts.

8. Conclusion

We analyzed a periodic review inventory policy that expedites the delivery times of open orders when inventory drops below a threshold. We developed a solution approach that is easy to implement and that can be used to solve problems of realistic size. For inventory systems with variable expediting cost and FCFS expediting, the solution is optimal. For inventory systems with fixed expediting cost or free expediting, our policy is generally not optimal but delivered close-to-optimal results in numerical experiments.

For the data of the equipment manufacturer that motivated our research, we quantified the cost savings that can be achieved by order expediting. Our results indicate that a substantial savings potential exists and that the magnitude of the savings remains substantial for reasonable variations of the demand rates and lead times. In general, the cost savings increase in the lead time, the back-order penalty cost, and the inventory holding cost. Naturally, they decrease in the expediting cost but stay in the double-digit percentage range for the set of expediting cost parameters that we analyzed.

Although the policy performed well in most of the settings that we analyzed, we do not expect it to perform equally well in situations with substantial fixed order cost, where optimal order quantities and optimal expediting quantities can be large. Solution approaches that are tailored to such settings are likely to perform better than our policy. However, they are also likely to be complex, and we leave their analysis to future research. We also leave the analysis of other extensions to future research, such as the analysis of stochastic lead times.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2014.0506.

References

Allen SG, D'Esopo DA (1968) An ordering policy for stock items when delivery can be expedited. *Oper. Res.* 16(4):880–883.

Barankin EW (1961) A delivery-lag inventory model with an emergency provision (the single-period case). *Naval Res. Logist.* 8(3):285–311.

Berling P, de Albeniz VM (2011) Optimal expediting decisions in a continuous-stage serial supply chain. Working paper, IESE Business School, Barcelona.

Bertsekas DP (2007) *Dynamic Programming and Optimal Control*, Vol. II, 1st ed. (Athena Scientific, Belmont, MA).

Bijvank M, Vis IFA (2011) Lost-sales inventory theory: A review. *Eur. J. Oper. Res.* 215(1):1–13.

Chiang C (2002) Ordering policies for inventory systems with expediting. *J. Oper. Res. Soc.* 53(11):1291–1295.

Chiang C (2010) An order expediting policy for continuous review systems with manufacturing lead-time. *Eur. J. Oper. Res.* 203(2):526–531.

Chiang C, Gutierrez GJ (1996) A periodic review inventory system with two supply modes. *Eur. J. Oper. Res.* 94(3):527–547.

Clark AJ, Scarf H (1960) Optimal policies for a multi-echelon inventory problem. *Management Sci.* 6(4):475–490.

Duran A, Gutierrez G, Zequeira R (2004) A continuous review inventory model with order expediting. *Internat. J. Production Econom.* 87(2):157–169.

Jain A, Groenevelt H, Rudi N (2010) Continuous review inventory model with dynamic choice of two freight modes with fixed costs. *Manufacturing Service Oper. Management* 12(1):120–139.

Kelle P, Silver EA (1990) Safety stock reduction by order splitting. *Naval Res. Logist.* 37(5):725–743.

Lawson DG, Porteus EL (2000) Multistage inventory management with expediting. *Oper. Res.* 48(6):878–893.

Minner S, Diks EB, De Kok AG (2003) A two-echelon inventory system with supply lead time flexibility. *IIE Trans.* 35(2):117–129.

Muharremoğlu A, Tsitsiklis JN (2003) Dynamic leadtime management in supply chains. Preprint, Massachusetts Institute of Technology, Cambridge.

Odoni AR (1969) On finding the maximal gain for Markov decision processes. *Oper. Res.* 17(5):857–860.

Sculli D, Wu SY (1981) Stock control with two suppliers and normal lead times. *J. Oper. Res. Soc.* 32(11):1003–1009.

Sedarage D, Fujiwara O, Luong HT (1999) Determining optimal order splitting and reorder level for *N*-supplier inventory systems. *Eur. J. Oper. Res.* 116(2):389–404.

Sheopuri A, Janakiraman G, Seshadri S (2010) New policies for the stochastic inventory control problem with two supply sources. *Oper. Res.* 58(3):734–745.

Song J-S, Zipkin P (2013) Supply streams. *Manufacturing Service Oper. Management* 15(3):444–457.

Tagaras G, Vlachos D (2001) A periodic review inventory system with emergency replenishments. *Management Sci.* 47(3):415–429.

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.

White DJ (1963) Dynamic programming, Markov chains, and the method of successive approximations. *J. Math. Anal. Appl.* 6(3):373–376.

Zhou D, Leung LC, Pierskalla WP (2011) Inventory management of platelets in hospitals: Optimal inventory policy for perishable products with regular and optional expedited replenishments. *Manufacturing Service Oper. Management* 13(4):420–438.

Zhou SX, Chao X (2010) Newsvendor bounds and heuristics for serial supply chains with regular and expedited shipping. *Naval Res. Logist.* 57(1):71–87.