Optimal production and rationing decisions in supply chains with information sharing

Boray Huang\textsuperscript{a}, Seyed M.R. Iravani\textsuperscript{b,}\textsuperscript{*}

\textsuperscript{a}School of Business Administration, University of Mississippi, University, MS 38677, USA
\textsuperscript{b}Department of Industrial Engineering and Management Sciences, Northwestern University, 2143 Sheridan Road, C210 Tech., Evanston, IL 60208, USA

Received 9 October 2004; accepted 30 October 2006
Available online 4 January 2007

Abstract

This paper considers a two-echelon capacitated supply chain with two non-identical retailers and information sharing. We characterize the optimal inventory policies. We also study the benefits of the optimal stock rationing policy over the first come first served (FCFS) and the modified echelon-stock rationing (MESR) policies.

\textcopyright 2006 Elsevier B.V. All rights reserved.

Keywords: Stock rationing; Information sharing; Markov decision process; Supply chain; Production scheduling

1. Introduction

Stock rationing has recently become a popular issue in supply chain management. As indicated in [17,27], stock rationing is especially important when applying the following strategies in a supply chain: component commonality, delayed differentiation and inventory centralization. Earlier literature on stock rationing assumes uncapacitated exogenous supply (see, for example, [26,10,13]). Ha [17] considers the problem of stock rationing in a capacitated make-to-stock production system with multi-class demand and lost sales. He shows that the threshold-type policies are optimal for both production and rationing decisions. Similar results are provided in [18] for systems with two demand classes and backlog. The model in [18] is extended by [27] for systems with \( n \) demand classes. They demonstrate that the benefit of inventory pooling can be realized only if the stock is efficiently allocated. For other extensions, see [12,16,2] and the references therein.

Most of the existing stock rationing papers in capacitated systems assume that the manufacturer faces customers’ unitary demands, so the rationing decision is in fact the decision whether to accept or reject an order. In supply chains, however, the manufacturer receives orders from its downstream retailers, who face
the end-customer demand. The downstream retailers’ order sizes are usually more than one. Under these circumstances, the problem of optimal rationing quantity arises, which corresponds to the question of what proportion of a retailer’s order must be satisfied. Huang and Iravani [21] consider a simple case in which a single-level make-to-stock system faces a compound Poisson demand process. They characterize the optimal production and stock rationing policies to be of the threshold type. This paper is different from [21], in the sense that: (i) this paper investigates the production/rationing policies in a two-echelon supply chain, and (ii) this paper considers the two-echelon supply chain under information sharing.

Information sharing becomes more and more critical in supply chain cooperation. Due to the rapid development of information technology such as EDI and the Internet, the participants in a supply chain can easily exchange their information regarding production, inventory and sales in a timely fashion. Through this shared information, businesses can gain a competitive advantage by developing more effective production and inventory control policies, which can bring higher profits and better customer service. The benefit of the information sharing has been extensively investigated in the existing literature (see [24,5] for excellent reviews). Most existing research on information sharing presumes either an uncapacitated exogenous supply (e.g., [3,8]), or if the supply is capacitated, the research assumes a single retailer (e.g., [15]), or multiple retailers for whom, due to equal shortfall/backlog cost for all retailers, a stock rationing decision is not required (e.g., [20]).

Our study differs from the existing literature on stock rationing or information sharing in the following three ways. First, we study the manufacturer’s optimal production/rationing policies in a two-level supply chain with non-identical retailers. Current research on stock rationing in capacitated systems usually focuses on a single manufacturer facing end-customer demand without any intermediate retailers. An exception is [2], which assumes that all the retailers apply base-stock policies. However, because the supplier still faces unitary orders, the retailers’ inventory information has no value to the supplier. In our paper the retailers order in batches; therefore, the manufacturer can utilize the shared inventory information to optimize its own production and rationing decisions.

Second, since the demand faced by the manufacturer is not unitary, the quantity of rationing becomes an interesting issue. Third, the manufacturer has information about downstream inventory levels. Therefore, the optimal stock rationing policy depends not only on the manufacturer’s on-hand inventory, but also on the retailers’ inventory levels.

This paper is organized as follows. In Section 2 we introduce our model and we formulate the production and stock rationing problem as a Markov decision process (MDP). In Section 3 the structural properties of the optimal policies are characterized. In Section 4 we introduce the modified echelon-stock rationing (MESR) policy and compare the performance of the optimal rationing policy with those of the first come first served (FCFS) and MESR policies.

2. Model description

We consider a two-level supply chain with one manufacturer and two retailers in a competitive market. The manufacturer produces items in the production facility and keeps them in its own inventory in order to satisfy the retailers’ orders. There is a convex, non-decreasing and non-negative holding cost $h(y)$ per unit time when the manufacturer’s inventory level is $y$. The manufacturer has a limited capacity of producing $\mu$ items per unit time. The end consumers arriving at Retailer $i$ have a unitary demand, which is independent of the demand at the other retailer, and follows a Poisson process with rate $\lambda_i$, $i = 1, 2$. Both retailers use $(Q_i, R_i)$ policies to manage their inventories ($i = 1, 2$).

That is, when Retailer $i$’s inventory level reduces to the reorder point $R_i$, an order in size of $Q_i$ units is issued to the manufacturer. When an order from Retailer $i$ arrives, the manufacturer has to decide how to satisfy the order from the manufacturer’s on-hand inventory. If the order in a size of $Q_i$ units is not entirely satisfied, the manufacturer incurs a linear shortfall penalty $c_i$ per unit for the unmet part of the order.

The shortfall penalty $c_i$ represents the loss of revenue or goodwill due to a shortfall (see [1]). In a competitive market, the manufacturers will usually try different ways to satisfy the unmet portion of the retailers’ orders. For example, (i) the manufacturer can purchase the unmet part from other suppliers,
including other strategically allied manufacturers in different geographic areas [15,23], or (ii) the manufacturer can expedite (e.g., use overtime) to produce the unmet portion of the order [11,14]. In both cases (i) and (ii), Retailer i will get his entire order of size \( Q_i \) from the manufacturer; however, the manufacturer faces cost \( c_i \) for every item obtained through purchasing or expediting. Without loss of generality, we assume Retailer 1’s orders have the larger shortfall penalty rate than Retailer 2’s, i.e., \( c_1 > c_2 \).

We assume that all this happens during an insignificant transportation lead-time. As a result, \( R_i = 0 \), for \( i = 1, 2 \), so that the retailers’ on-hand inventory levels are always between 1 and \( Q_i \). Because the transportation lead-times affect only the retailers’ reorder points in a decentralized supply chain, it is not too difficult to extend our analysis of the manufacturer’s policies to cases with positive transportation lead-times. With the objective of minimizing its total cost, the manufacturer faces two kinds of decisions: production control decisions and inventory allocation decisions.

Assuming that production times follow exponential distribution, we formulate the manufacturer’s production and rationing decisions as an MDP. The MDP is defined as follows:

- The state space \( \mathcal{S} \) consists of the triplet \( (y, x_1, x_2) \), where \( y \) and \( x_i \) are the manufacturer’s and Retailer \( i \)'s on-hand inventory, respectively. Thus,

\[
\mathcal{S} = \{(y, x_1, x_2) | y, x_1, x_2 \in \mathbb{Z}^+, Q_1 \geq x_1 \geq 1, Q_2 \geq x_2 \geq 1\},
\]

where \( \mathbb{Z}^+ \) is the set of non-negative integer numbers.

- Decision epochs are the end-customer demand arrival epochs at the retailers or the production completion epochs at the manufacturer’s production facility.

- The action space includes three actions: (i) idling, (ii) production and (iii) rationing \( z \) items from the manufacturer’s on-hand inventory when an order comes from Retailer \( i \) (\( \min \{Q_i, y\} \geq z \geq 0, z \in \mathbb{Z}^+, i = 1, 2 \)).

We note here that the exponential assumption regarding the production times allows us to formulate this MDP and derive an optimal policy. After our MDP reveals the structure of the optimal production and rationing policies, it becomes clear that our main insights are not influenced by the assumption on production times.

When an order from Retailer 1 comes, the fact that \( c_1 > c_2 \) makes it obvious that the manufacturer’s optimal decision is to fill the order as completely as possible. That is, the rationing problem arises only when an order from Retailer 2 arrives. Without loss of generality, we scale the parameters and let \( \mu + \lambda_1 + \lambda_2 + x = 1 \). The optimality equation for the MDP to minimize the total discounted holding and shortfall cost is:

\[
V_2(y, x_1, x_2) = h(y) + \lambda_1 V_2[(y - 1) Q_1^+, x_1 - 1] + \lambda_2 V_2[(1 - \lambda_2) V_2(y, x_1, x_2 - 1)] + \mu \min \{V_2(y - z, x_1, Q_2) + c_2 [Q_2 - z] \}
\]

where \( \min \{Q_2, y\} \geq z \geq 0 \), and \( V_2(y, x_1, x_2) \) is the optimal discounted cost under initial state \((y, x_1, x_2)\). Furthermore,

\[
[X] = \begin{cases} X & \text{if } X > 0, \\ 0 & \text{if } X \leq 0. \end{cases}
\]

The last two lines on the right-hand side of (1) represent the manufacturer’s optimal rationing and production decisions. When an order from Retailer 2 arrives (where \( \lambda_2 = 1 \)), the manufacturer’s best decision is to allocate \( z_{y,x_1}^* \) items from its on-hand inventory. The manufacturer is then charged the shortfall cost \( c_2 \) per unit for the remaining part of the order. Thus, considering \( \min \{Q_2, y\} = Q_2 - [Q_2 - y]^+ \), under the discounted-cost optimality, we will have

\[
z_{y,x_1}^* = \arg \min_{0 \leq z \leq Q_2 - [Q_2 - y]^+} \{V_2(y - z, x_1, Q_2) + c_2 [Q_2 - z]\}.
\]

The last term on the right-hand side of (1) represents the manufacturer’s optimal production decisions. If \( V_2(y, x_1, x_2) \leq V_2(y + 1, x_1, x_2) \), idleness is optimal at state \((y, x_1, x_2)\); otherwise, production is optimal. We let idleness be the optimal decision when an equality occurs. Note that the average-cost optimality can be obtained by letting the discount factor \( \mu \) approach zero (see [28,19]). Thus, in the following section we
provide the structural analysis of the optimal policies only for the discounted-cost criterion.

3. Characteristics of the optimal policies

In this section we analyze the properties of the optimal policies. The proofs for convexity and supermodularity of the cost function \( V_z \) as well as the existence of the optimal stationary policy and the structural results can be found in [22], the full version of this paper.

**Theorem 1 (Optimal production policy).** The manufacturer’s optimal production policy has the following properties:

(a) State-dependent modified base-stock policies are optimal for the manufacturer’s production decision.

(b) The manufacturer’s optimal base-stock levels are non-increasing in Retailer 1’s on-hand inventory \( x_1 \).

(c) If idleness is optimal at state \((y, x_1, x_2)\), then regardless of Retailer 1’s on-hand inventory level, idleness is also optimal when the manufacturer’s and Retailer 2’s on-hand inventory levels are at least \( y + Q_1 \) and \( x_2 \), respectively. Therefore, for any given \( x_2 \), the difference between the highest and the lowest base-stock levels is no greater than \( Q_1 \).

**Theorem 2 (Optimal rationing policy).** The manufacturer’s optimal stock rationing policy has the following properties:

(a) State-dependent reserve-stock policies are optimal for the manufacturer’s rationing decision. There exist reserve-stock levels \( r_{x_1}^* \) such that the manufacturer’s optimal rationing decision is to make the manufacturer’s inventory level after the allocation as close as possible to the corresponding reserve-stock level. That is,

\[
z^*_{y,x_1} = \min \{ Q_2, \lfloor y - r_{x_1}^* \rfloor^+ \},
\]

where

\[
r_{x_1}^* = \min \{ y \mid V_2(y + 1, x_1, Q_2) - V_2(y, x_1, Q_2) \geq -c_2, \ y \in \mathbb{Z}^+ \}.
\]

(b) The optimal reserve-stock levels \( \{r_{x_1}^*\} \) are non-increasing in Retailer 1’s on-hand inventory \( x_1 \).

(c) The difference between the highest and the lowest optimal reserve-stock levels is no greater than \( Q_1 \) (i.e., \( r_{x_1}^* - r_{Q_1}^* \leq Q_1 \)).

According to Theorem 2, when an order from Retailer 2 arrives, the manufacturer should reject the whole order when its inventory is no more than \( r_{x_1}^* \). On the other hand, when the manufacturer has more than \( r_{x_1}^* \) items on hand, it should allocate as many items as possible (up to \( Q_2 \)) to fill the order as long as its inventory level after the allocation does not go below \( r_{x_1}^* \). Note that \( r_{x_1}^* \) does not depend on the manufacturer’s on-hand inventory \( y \).

Fig. 1 left shows the relation between the inventory levels before and after the stock rationing under the optimal policy. The horizontal axis represents the inventory level \( y' \) before the allocation, and the vertical axis represents the inventory level \( y'' \) after the allocation. Note that the difference between the bold line and the dotted line shows the optimal rationing amount \( z^*_{y,x_1} \).

Fig. 1 right shows an example of the optimal production/inventory policy and the optimal rationing policy at \( x_2 = 1 \) when \( \mu = 1.5, z_1 = z_2 = 1.0, Q_1 = 15, Q_2 = 10, c_1 = 30 \) and \( c_2 = 10 \). Because Retailer 2 has only one item on hand \( (x_2 = 1) \), the stock rationing problem arises if a customer arrives at Retailer 2. The solid line in Fig. 1 right represents the base-stock levels and the dotted line represents the reserve-stock levels \( r_{x_1}^* \) for rationing decisions. According to Theorem 1, both lines are non-increasing with Retailer 1’s inventory level \( x_1 \). The highest base-stock level is 25 and the lowest is 11. The highest reserve-stock level is 15 and the lowest is zero. We illustrate the optimal decisions on three different points when Retailer 1 has 6 items on hand (i.e., \( x_1 = 6 \)), and when the manufacturer’s inventory is: \( A : \{ y = 25 \}, B : \{ y = 15 \} \) and \( C : \{ y = 6 \} \). When the system state is at point \( A \), the manufacturer should stay idle. If an order from Retailer 2 arrives, the manufacturer should fill the entire order from its on-hand inventory. At point \( B \) the optimal decision is to produce. If Retailer 2’s order in size of \( Q_2 = 10 \) arrives, the manufacturer should allocate only 7 items because she needs to keep 8 items for future orders \( r_{x_1}^* = 8 \). If the manufacturer has only 6 items on hand (point \( C \), the optimal
These expected costs can be obtained through the TC icy. To avoid the impact of the discount factor due to implementing the optimal stock rationing policy, the decision is to keep producing and reject any order from Retailer 2.

4. Cost-effectiveness of optimal stock rationing

In this section we perform an extensive numerical study to provide insights into the cost saving due to implementing the optimal stock rationing policy. To avoid the impact of the discount factor, we evaluate the performance of the optimal stock rationing policy and the benchmark policies by obtaining their long-run total average costs per unit time. We consider two benchmark policies: the FCFS policy and the MESR policy, which we introduce in Section 4.2. More specifically, we compare the manufacturer’s expected inventory cost under the optimal rationing policy, $TC_{opt}$, with those under the FCFS policy, $TC_{FCFS}$, and the MESR policy, $TC_{MESR}$. These expected costs can be obtained through the successive approximation with an error bound 0.01% (see [25] for the convergence of the successive approximation method). Note that $TC_{opt}$, $TC_{FCFS}$ and $TC_{MESR}$ are the expected costs under the optimal, the FCFS rationing and the MESR rationing policies with their corresponding optimal production policies, respectively.

Our numerical study consists of 5120 cases that were generated by considering different values of: (i) order sizes $Q_i \in \{1, 5, 10, 30\}$, $i = 1, 2$; (ii) demand rate ratio $\lambda_1/\lambda_2 \in \{0.05, 0.25, 1, 4, 20\}$; (iii) traffic intensity $\rho = (\lambda_1 + \lambda_2)/\mu$, where $\rho \in \{0.4, 1.0, 1.8, 4.0\}$; (iv) lost sales ratio $\beta = c_1/c_2$, where $\beta \in \{1.2, 2.5, 10\}$ and (v) relative holding cost rate $h' = h/(\lambda_1 c_1 + \lambda_2 c_2)$, where $h' \in \{0.002, 0.01, 0.02, 0.1\}$. For each case of our numerical study we calculate the relative cost reduction (cost saving) $CR$ in the following way:

$$CR_\pi = \frac{TC_\pi - TC_{opt}}{TC_\pi} \times 100\%, \quad \pi = FCFS, MESR.$$  

The quantity $CR$ measures the percent cost saving due to implementing the optimal stock rationing policy instead of the benchmark policy $\pi$. The larger $CR$ is, the more beneficial it will be to employ the optimal rationing policy.

4.1. Optimal rationing versus FCFS

Based on the numerical examples we observed that, for given order quantities, the impacts of the demand rate ratio, traffic intensity and relative holding cost rates on system performances were similar to those of unitary-demand systems in [17]. Therefore, we only present the impact of order sizes $\{Q_i\}$, $i = 1, 2$, on the manufacturer’s total average cost and on the benefit of the optimal rationing policy. Note that under unitary orders, i.e., $Q_1 = Q_2 = 1$, the analysis of our supply chain system is the same as the analysis of the make-to-stock system presented in [17].

Fig. 2 shows the impact of order sizes on the benefit of stock rationing $CR_{FCFS}$, the base-stock level and the reserve-stock level in one of our cases where $\lambda_1/\lambda_2 = 1.0$, $\beta = 5.0$, $\rho = 1.8$ and $h' = 0.01$. We
increase one of the order sizes \( \{Q_i\} \) \((i = 1, 2)\) from 1 to 30, while keeping the other order size \( Q_j (j \neq i)\) fixed. Fig. 2 right shows the optimal base-stock levels (S) and the reserve-stock levels (r) for a case when both retailers have only one unit in their inventories (i.e., when \( x_1 = x_2 = 1 \)). From Fig. 2 we observe that:

- When Retailer 1 orders less frequently but with a larger order size, both \( T_{opt} \) and \( T_{FCFS} \) increase. \( CR_{FCFS} \) also increases with \( Q_1 \), especially when \( Q_2 \) is large (see Fig. 2 left). This implies that, in a supply chain with information sharing, batch ordering from the more valuable retailers makes the optimal rationing policy more cost-effective than that in the unitary-demand model of [17]. The result is also different from [21], in which \( CR_{FCFS} \) dramatically decreases with customer’s order sizes. It is due to the fact that, under the optimal stock rationing policy, the manufacturer has to reserve more stock for a large \( Q_1 \) to avoid high shortfall penalty from Retailer 1 (see Fig. 2 right). When \( Q_1 \) is large and the retailers’ inventory information is shared, the manufacturer can better schedule its production, shorten the holding time of reserved stock and thus increase the efficiency of stock rationing. As a result, the benefit of optimal stock rationing can remain significant when \( Q_1 \) is large in our information sharing supply chain.

- When Retailer 2 orders less frequently but with a larger order size, \( CR_{FCFS} \) decreases significantly with \( Q_2 \) (see Fig. 2 left). This is because, similar to [21] for a one-level production system, the manufacturer can enjoy a lower total cost \( T_{FCFS} \) and even a lower base-stock level if the FCFS rationing policy is used and the less valuable retailer has a larger order size. As a result, the optimal stock rationing does not enjoy much more advantage over the FCFS rationing when \( Q_2 \) increases. Furthermore, Ha [17] reports that the maximal benefit of optimal rationing in a make-to-stock system with unitary demand can be up to 60%. In our 5120 examples, we found that \( CR_{FCFS} \) can be up to 80% in our supply chain setting with information sharing, with an average-cost reduction 7.2%. Since the retailers’ order sizes can significantly change the benefit of optimal stock rationing, managers may need to consider different strategies to induce proper order quantities from different downstream retailers when the optimal stock rationing is applied.

4.2. Optimal rationing versus MESR

In this section we introduce the MESR policy and compare its performance with the performance of the optimal stock rationing policy. The MESR policy uses a threshold value \( T \) as the basis for making rationing decisions.
MESR Policy: Upon arrival of an order of Retailer 2, if the sum of the manufacturer’s inventory \( y \) and Retailer 1’s inventory \( x_1 \) is less than a threshold \( T \), i.e., \( y + x_1 < T \), then the manufacturer will reject Retailer 2’s entire order. Otherwise, if \( y + x_1 > T \), then the manufacturer will try to fill Retailer 2’s order so that the echelon’s postallocation stock level does not fall below \( T \). More specifically, upon Retailer 2’s order arrival at state \((y, x_1, x_2)\), the manufacturer will allocate \( a_{y,x_1} \) units from its on-hand inventory to fill Retailer 2’s order, where

\[
a_{y,x_1} = \begin{cases} 
0 & \text{if } y + x_1 < T, \\
\min\{y + x_1 - T, y, Q_2\} & \text{if } y + x_1 > T.
\end{cases}
\]

Compared with the optimal policy, in which the reserve-stock levels vary with \( x_1 \), the MESR policy uses only one threshold value to make the rationing decision. Therefore, it is easy to implement in practice. It is important to also note that when \( Q_1 = Q_2 = 1 \), the optimal stock rationing policy and the MESR policy become identical.

Echelon-stock policies are generally used to make production decisions (see [9,4,6,7]). Here we modified its structure so it can be used to make rationing decisions. We did not use the echelon-stock policy as the production policy in our study, since it has been shown in [20] that for cases with \( c_1 = c_2 \) the echelon-stock policies are not cost-effective policies for the manufacturer’s production decision in a multi-retailer supply chain with information sharing.

For each case of our numerical study we searched for the optimal threshold \( T^* \) and used its corresponding cost \( TC_{MESR} \) to calculate \( CR_{MESR} \). We found that, in general, the MESR policy performs close to the optimal stock rationing policy. The average value of \( CR_{MESR} \) in our examples is around 2%. However, we found that, the optimal stock rationing policy is significantly more cost-effective than the MESR policy when \( \lambda_1/\lambda_2 \) is small, \( Q_1/Q_2 \) is large and \( \beta \) is large. (See Fig. 3 left for an illustration when \( Q_2 = 1, \rho = 1.0 \) and \( h' = 0.01 \).) Under these circumstances, using the optimal stock rationing policy instead of the MESR policy can save up to 28% in overall cost.

Fig. 3 right is another representative example that depicts the impact of the order size \( Q_1 \) on \( CR_{MESR} \) when \( \lambda_1/\lambda_2 = 0.05, \rho = 1.0, \beta = 10, h' = 0.01 \). As the figure shows, as \( Q_1 \) increases, \( CR_{MESR} \) increases. The reason is as follows: since \( Q_1 \) and \( \beta \) are large, it is better for the manufacturer to reserve more stock when \( x_1 \) is low. Under the MESR policy, this means a higher value of \( T^* \), and the manufacturer may need to reject Retailer 2’s orders even when \( x_1 \) is high. For example, suppose \( Q_1 = 30 \). Under the MESR policy with \( T^* = 31 \), the manufacturer should try to keep at least 30 units on hand when a Retailer 2 order arrives and \( x_1 = 1 \). When a Retailer 2 order arrives and \( x_1 = 25 \), the manufacturer still has to reject the entire order as long as its inventory level is no more than 6. On the other hand, the optimal stock rationing policy can take advantage of relatively small \( \lambda_1 \), which implies
that the manufacturer does not need to reserve stock until Retailer 1’s inventory level is very close to the bottom. That is, \( r_1^* \) can stay zero until \( x_1 < 5 \), and dramatically increase to 30 when \( x_1 = 1 \). Compared with the MESR policy, the optimal stock rationing can therefore save more in inventory holding cost and in the penalty cost of Retailer 2’s orders. As a result, the optimal rationing policy becomes more cost-effective than the MESR rationing policy in cases when \( \lambda_1/\lambda_2 \) are small, \( \beta \) and \( Q_1/Q_2 \) are large.

References

[22] B. Huang, S. Iravani, Optimal production and rationing decisions in supply chains with information sharing, Working paper, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, 2006 (full version of this paper).