M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 16, No. 4, Fall 2014, pp. 561–577
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2014.0494

Exact Analysis of Capacitated Two-Echelon Inventory Systems with Priorities

Hossein Abouee-Mehrizi
Department of Management Sciences, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada,
haboueemehrizi@uwaterloo.ca

Opher Baron, Oded Berman
Joseph L. Rotman School of Management, University of Toronto, Toronto M5S 3E6, Canada
{opher.baron@rotman.utoronto.ca, berman@rotman.utoronto.ca}

We consider a two-echelon inventory system with a capacitated centralized production facility and several distribution centers (DCs). Both production and transportation times are stochastic with general distributions. Demand arrives at each DC according to an independent Poisson process and is backlogged if the DC is out of stock. We allow different holding and backlog costs at the different DCs. We assume that inventory at DCs is managed using the one-for-one replenishment policy. The main objective of this paper is to investigate the control of the multiechelon $M/G/1$ setting with general transportation times. To achieve this objective, we analyze several decentralized allocation policies including the first-come, first-served (FCFS), strict priority (SP), and multilevel rationing (MR) policies. For our analytic results, we assume no order crossing. We derive the cost function for a capacitated two-echelon inventory system with general transportation times under these policies. Our numerical examples show that the FCFS policy may outperform the MR policy, even though the latter has been shown to be better in the centralized setting. This suggests that in decentralized settings there is a need to focus on policies that prioritize customers when there is backlog. This focus is in contrast to the centralized settings, where inventory rationing policies that focus on prioritization when there is available inventory are effective. We therefore introduce and analyze the generalized multilevel rationing (GMR) priority policy. We compare the GMR policy with other policies and show that the GMR policy outperforms the three policies used in the centralized setting. We also compare the GMR policy with the myopic (T), longest queue first (LQF), and the optimal (when order crossing is allowed during the transportation time) policies. Our results show that when the uncertainty of the transportation times is low, the GMR policy outperforms the myopic (T) and LQF policies and that the gap between the optimal policy and the GMR policy is not high.

Keywords: multiechelon inventory management; $M/G/1$ queue; strict priority; multilevel rationing; stochastic production times; stochastic transportation times
History: Received: September 5, 2012; accepted: April 26, 2014. Published online in *Articles in Advance* August 21, 2014.

1. Introduction

Inventory management is an important part of supply chain operations. An IBM report in 2009 indicates that a 30%–60% reduction in spare parts inventory at 30 distribution centers (DCs) can save up to US$500 million per year (Mak and Shen 2009). One common strategy to manage inventory and reduce the inventory cost is customer prioritization. For example, the empirical study by Jing and Lewis (2011) shows that increasing the service level of new customers to 95% by keeping some inventory on hand for them can reduce the long-term stockout cost by up to 8.4%. Their study highlights the impact of a priority policy on the long-term economic value of different customer segments.

Although the potential benefits of customer prioritization have been pointed out in the literature, this paper is the first to investigate the possible benefit in the capacitated multiechelon inventory management setting. Specifically, we study a two-echelon inventory system with a capacitated production facility that keeps inventory at the warehouse and satisfies the demand of several DCs. Inventory at the warehouse is replenished from a production facility. We allow the production times to be generally distributed. Demand arrives at each DC according to a Poisson process and demand that is not satisfied immediately from on-hand inventory is backlogged. DCs keep stock to reduce the stockout cost and replenish their inventory from the warehouse using the one-for-one replenishment policy. We allow the transportation times between the DCs and the warehouse to have a general distribution. For this decentralized inventory management problem, our objective is to minimize the total holding and backlog costs of the system by allowing customer prioritization at the warehouse.

Our model with the assumptions of Poisson demand, continuous review, and one-for-one inventory replenishment policy can be used for spare parts inventory management (see, e.g., Axsater 2006). For example, Schneeweiss and Schroder (1992) consider a hierarchical model for Lufthansa to manage the spare parts inventory. In their model, similar to ours, order arrivals follow a Poisson process and the inventory is managed using the one-for-one replenishment policy. They state that the implementation of their model by Deutsche Lufthansa AG provided substantial savings. As another example, consider an iPad that has failed under warranty. In such a case, customers make an appointment at one of the Apple stores, and after demonstrating that the failed item is under warranty they receive a refurbished iPad. The malfunctioning item is sent to the refurbishing facility (representing a new demand that can be modeled as coming from a Poisson process), where it will be repaired and used to satisfy a future demand. As the iPad is sold in many countries, both the transportation time from the refurbishing facility to the different stores and the prices paid at these stores may vary. Thus, customer prioritization may be valuable and our model may be applicable in supply chains of high-tech products under warranty.

1.1. Main Contributions

We consider a two-echelon inventory system with a capacitated manufacturer and multiple DCs. The production and transportation times are generally distributed and demand arrivals follow Poisson processes. For our analytical results, we assume no order crossing during the transportation from the warehouse to DCs. Using the queueing decomposition (QD) approach introduced in Abouee-Mehrizi et al. (2012; hereafter, ABB) and further explained in Abouee-Mehrizi and Baron (2014), we first obtain the production plus waiting time of an order at the warehouse under the first-come, first-served (FCFS), strict priority (SP), and multilevel rationing (MR) policies. Then, following Svoronos and Zipkin (1991), we derive the distribution of backlog and inventory levels at each echelon to characterize the total cost of each DC.

We next compare these policies numerically when their prioritization is based on the backlog costs (i.e., a DC with a higher backlog cost is given a higher priority). Although ABB analytically show that the MR policy outperforms the SP and FCFS policies in the centralized (single-echelon) $M/G/1$ settings, we find that this is not necessarily the case in the multiechelon inventory systems. Our numerical results demonstrate that although the total cost of the system under SP and MR policies is often lower than under the FCFS policy, there are cases where the FCFS policy results in a lower cost. These results indicate that the priority policies that have been considered for centralized inventory systems are less efficient in the multiechelon setting. We thus consider a new priority policy for the multiechelon inventory systems, namely, the general multilevel rationing (GMR) policy. Unlike the MR policy, the GMR policy allows a DC to be prioritized when the inventory level hits a negative threshold, i.e., when the number of backlogs at the warehouse hits a threshold. We show that the FCFS, SP, and MR policies are special cases of this policy.

The four policies discussed above use no information about the inventory in transit or at the DCs. That is, these policies only use local information at the level of warehouse inventory and backlog. Limiting the information simplifies their implementation but renders them not optimal in general.

In general, when transportation times are stochastic, orders may cross each other. To better evaluate the performance of the GMR, we also consider cases with order crossing. We obtain the total cost of the system under the longest queue first (LQF), myopic (T), and GMR with order crossing (GMROC) policies using simulation, and evaluate the performances of these policies compared to the optimal policy (found using dynamic programming). We demonstrate that order crossing can significantly impact the total cost of the system.

The main contributions of this paper are threefold. (1) We derive the optimal cost and base-stock levels in a multiechelon inventory system with a capacitated manufacturer where the stock allocation at the manufacturer follows the FCFS policy. As mentioned earlier, although there is a vast body of literature considering the FCFS policy in capacitated systems, this is the first paper to provide an exact analysis of a multiechelon inventory system under this policy. (2) We investigate customer prioritization in a two-echelon inventory system with a capacitated production facility. As part of this investigation, we introduce and analyze the GMR priority policy, derive its optimal cost and control, and compare it to other policies. Furthermore, we provide an extensive numerical study comparing the performances of six inventory allocation policies (with and without order crossing), highlighting the potential benefits of such policies under different settings. For example, in our numerical study, the GMR policy outperforms the FCFS policy by 9% on average. (3) We suggest an additional application of QD, namely, that it can be used to derive the exact optimal solution for several priority policies in multiechelon $M/G/1$ make-to-stock systems with general transportation times (and no order crossing).

1.2. Literature Review

For a centralized inventory system, two priority policies based on the backlog costs have been considered in

the literature: the SP and MR policies. (We explain these policies in the next section.) Ha (1997a) is the first to consider the MR policy for a centralized single product multiclass make-to-stock system where the manufacturer has finite capacity and unsatisfied demand is backlogged. He shows that a policy with monotone switching curves is optimal for a system with two classes of customers where the arrival process follows a Poisson process and production times are exponentially distributed. De Véricourt et al. (2002) extend the previous work to a system with multiple customer classes and show that the MR policy is optimal. De Véricourt et al. (2001) introduce the SP policy, compare the FCFS, SP, and MR policies for an $M/M/1$ make-to-stock queueing system, and show that the MR policy outperforms the other two policies. Benjaafar et al. (2010) investigate an $M/M/1$ make-to-stock system with two classes of customers with both backlogging and lost sales. They show that the optimal policy can be described by three state-dependent thresholds: a production base-stock level and two order-admission levels, one for each class. As mentioned earlier, ABB introduce the QD approach (they called it customer composition) to generalize the results for the $M/G/1$ make-to-stock systems with backlogging.

In the decentralized settings where a supplier with limited capacity satisfies demand for different types of products, dynamic scheduling policies have been considered. For example, Wein (1992) considers $b\mu/h\mu$ policy: when there are backlogs, prioritize the class with the higher backlog cost; if no classes are in danger of being backlogged, prioritize the class with the lower holding cost. Veatch and Wein (1994) suggest a kanban-like policy that works well for systems in which the second stage is the bottleneck. But in our model, the second stage (i.e., transportation time) is not the bottleneck of the system; thus, the efficiency of such policies is limited. Zheng and Zipkin (1990) analyze an $M/M/1$ make-to-stock system with two identical products under the longest queue first policy and compare its performance with the FCFS policy. Peña Perez and Zipkin (1997) propose another heuristic called myopic (T). Ha (1997b) and de Véricourt et al. (2000) characterize several structural properties of the optimal policy for such settings.

Multiechelon inventory systems with ample supply have been studied since the seminal work by Clark and Scarf (1960). For an uncapacitated serial inventory system, they characterize the optimal policy for a finite-horizon setting. Sherbrooke (1968) introduces the METRIC model to approximate the number of backlogs at each echelon for a two-echelon inventory system with one-for-one replenishment policy. Graves (1985) approximates the total cost of a similar system with a central warehouse and several DCs where the lead time (transportation time) at the warehouse is generally distributed and the transportation times are deterministic. Svoronos and Zipkin (1991) evaluate a one-for-one replenishment policy in a serial system setting, showing that the distribution of backlog and inventory level at each echelon can be obtained using the convolution of lead times at each echelon. Muharremoglu and Tsitsiklis (2008) consider an uncapacitated serial inventory system and characterize the optimal policy. Levi et al. (2008) study a two-echelon inventory system with one warehouse and multiple DCs and develop an algorithm to approximate the optimal base-stock levels at each echelon.

Capacitated serial multiechelon inventory systems have also been studied in the literature. Glasserman (1997) studies a multiechelon periodic-review inventory model where the supplier has limited capacity in each period and provides bounds for the optimal base-stock levels for a serial system with an echelon base-stock policy. Roundy and Muckstadt (2000) consider a similar problem and present an efficient approximation for the distribution of the shortfall process in each period. Parker and Kapuscinski (2004) show that a modified echelon base-stock policy is optimal for a capacitated two-echelon serial inventory system.

Few papers consider a capacitated nonserial mutliechelon inventory system that faces congestion. An exception is Mak and Shen (2009) who consider a two-echelon inventory system with a manufacturer operating from a warehouse and satisfying the demand arriving from several DCs. They assume demand at each DC follows a Poisson process, production times are exponentially distributed, transportation times are deterministic, and DCs are served based on a FCFS policy. They approximate the optimal base-stock levels at the DCs. Recently, Abouee-Mehrizi et al. (2011) use the unit-flow approach to obtain the optimal base-stock levels at the DCs for such a system.

1.3. Organization

The rest of this paper is organized as follows. In §2, we explain the capacitated two-echelon supply chain model in more detail. We discuss stock allocation polices in §3. In §4, we investigate the total cost at the DCs for given rationing and base-stock levels at the warehouse. Since the derivation of the optimal cost under the GMR policy is intricate, we analyze a system with two DCs. We analyze the cost of the warehouse and obtain the optimal rationing and base-stock levels in §5, and compare the cost of the different polices and derive some managerial insights in §6. We conclude the paper in §7. All proofs appear in Online Appendix A, and relevant results from ABB are provided in Online Appendix B. (Online appendices available as supplemental material at http://dx.doi.org/10.1287/msom.2014.0494.)

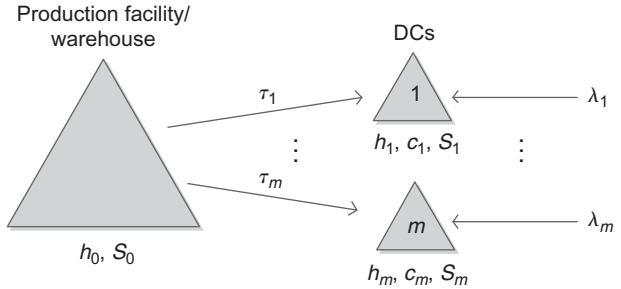
2. Capacitated Two-Echelon Supply Chain with Priorities

We study a supply chain with a production facility operating from a warehouse and m DCs as shown in Figure 1. For notational convenience, we number the warehouse as DC 0. We assume that the centralized warehouse is filled from the capacitated production facility with production times, χ, that follow a general distribution with a Laplace transform (LT) $\tilde{b}(\cdot)$ and a mean of $1/\mu$. Demand arrives at DC j according to a Poisson process with rate λ_j. We denote the total arrivals to the system by $\lambda := \sum_{j=1}^{m} \lambda_j$. For the stability of the system, we assume $\rho := \lambda/\mu < 1$. DC j may hold some stock with a holding cost of h_j per unit of inventory per unit of time. A customer arriving at DC j is served immediately if the inventory level is positive; otherwise, the order is backlogged with a backlog cost of c_j per unit of time. We denote the base-stock level at DC j by S_j and assume this inventory is replenished using a one-for-one policy. This means that DC j orders a new unit of the product from the warehouse as soon as a demand occurs at this DC.

With S_0 denoting the base-stock level at the warehouse, the production facility starts producing a new product as soon as the inventory level at the warehouse decreases to $S_0 - 1$. We assume that information on demand is immediately communicated to the warehouse; thus, such a decrease occurs whenever a demand arrives at either of the DCs. The production continues until the inventory level at the warehouse increases to S_0. When production ends, the product is either kept at the warehouse or sent to a DC depending on the prioritization policy and the inventory level at the warehouse. We discuss these issues in more detail in the following sections.

With these assumptions, the production facility can be modeled as an $M/G/1$ make-to-stock system with a centralized inventory location with postponement of the allocation decision until the end of the production. Since the backlog and holding costs of the DCs are different, prioritizing DCs at the warehouse and keeping some inventory for arrivals with higher backlogs or holding costs may reduce the total cost of the system.

Figure 1 Two-Echelon Inventory System



Note. h_j and S_j denote the holding costs rates and base-stock levels, respectively, at the warehouse and DCs $j = 0, \ldots, m$; c_j, τ_j, and λ_j denote the backlog costs, transportation times, and demand rates, respectively, at DCs $j = 1, \ldots, m$.

The transportation time between the warehouse and DC j, τ_j, is positive. We assume that the equilibrium transportation time between the warehouse and DC j, is generally distributed with a LT of $\tilde{L}_j(\cdot)$ and a mean of $1/\theta_j$. Moreover, we assume that DCs receive products from the warehouse sequentially, i.e., no order crossing in time is allowed and the queue discipline at each DC is FCFS. This is in contrast to the assumption of independent and identically distributed transportation times. (For a similar assumption see, e.g., Svoronos and Zipkin 1991.) We do not allow transshipment of inventory between the DCs.

Our objective is to determine the optimal rationing and base-stock levels at the warehouse and DCs to minimize the total expected holding and backlog costs. We assume that the allocation policy at the warehouse uses no information on the actual inventory level at the DCs. Since the demand arrival at DC j is independent of the arrivals at the other DCs, and the inventory at the DCs is managed using the one-for-one replenishment policy, the total cost of DC j is independent of the base-stock levels at the other DCs (see, e.g., Axsater 1990).

For notational convenience, we use superscript "$\bullet$" to denote policy "$\bullet$": FCFS, SP, MR, and GMR. Let $C_j^{\bullet}(S_0, S_j)$ denote the expected total holding and backlog costs at DC j, given policy $\bullet$ when the base-stock level at the warehouse is S_0 and the base-stock level at DC j is S_j. Similarly, let $C_0^{\bullet}(S_0)$ denote the expected holding cost at the warehouse. Note that the backlog cost of the system is captured in the DCs' cost functions, and therefore it is not included in the warehouse's cost function. The objective function of policy $\bullet$ can be expressed as

$$\text{Minimize } Z^{\bullet} = \sum_{j=1}^{m} C_j^{\bullet}(S_0, S_j) + C_0^{\bullet}(S_0). \tag{1}$$

Note that under the FCFS and SP polices, the cost function of DC j, $C_j^{\bullet}(S_0, S_j)$ depends only on the base-stock levels at the warehouse and DC j; however, under the MR and GMR polices, this cost also depends on the rationing levels at the warehouse. The expected holding cost at the warehouse also depends on the rationing levels under the MR and GMR policies. We describe these controls in more detail in the next section.

3. Stock Allocation Policies

In this section, we discuss stock allocation policies. Our focus is on analytically deriving the optimal cost for each of these easy to implement policies rather than characterizing the optimal control policy for this multiechelon system. Recall that we are considering a two-echelon inventory system with a production

Table 1 Stock Allocation Under the FCFS, SP, and MR Policies

		FCFS						SP						MR					
Event	E	I_0	B	B_1	B_2	I_1	I_2	I_0	B	B_1	B_2	I_1	I_2	I_0	B	B_1	B_2	I_1	I_2
1		3	{ }	0	0	1	1	3	{ }	0	0	1	1	3	{ }	0	0	1	1
2	1	2	{ }	0	0	1	1	2	{ }	0	0	1	1	2	{ }	0	0	1	1
3	1	1	{ }	0	0	1	1	1	{ }	0	0	1	1	1	{ }	0	0	1	1
4	2	0	{ }	0	0	1	1	0	{ }	0	0	1	1	1	{2}	0	1	1	0
5	2	0	{2}	0	1	1	0	0	{2}	0	1	1	0	1	{2, 2}	0	2	1	−1
6	1	0	{2, 1}	1	1	0	0	0	{2, 1}	1	1	0	0	0	{2, 2}	0	2	1	−1
7	2	0	{2, 1, 2}	1	2	0	−1	0	{2, 1, 2}	1	2	0	−1	0	{2, 2, 2}	0	3	1	−2
8	2	0	{2, 1, 2, 2}	1	3	0	−2	0	{2, 1, 2, 2}	1	3	0	−2	0	{2, 2, 2, 2}	0	4	1	−3
9	1	0	{2, 1, 2, 2, 1}	2	3	−1	−2	0	{2, 1, 2, 2, 1}	2	3	−1	−2	0	{2, 2, 2, 2, 1}	1	4	0	−3
10	1	0	{2, 1, 2, 2, 1, 1}	3	3	−2	−2	0	{2, 1, 2, 2, 1, 1}	3	3	−2	−2	0	{2, 2, 2, 2, 1, 1}	2	4	−1	−3
11	0	0	{1, 2, 2, 1, 1}	3	2	−2	−1	0	{2, 2, 2, 1, 1}	2	3	−1	−2	0	{2, 2, 2, 2, 1}	1	4	0	−3
12	0	0	{2, 2, 1, 1}	2	2	−1	−1	0	{2, 2, 2, 1}	1	3	0	−2	0	{2, 2, 2, 2}	0	4	1	−3
13	0	0	{2, 1, 1}	2	1	−1	0	0	{2, 2, 2}	0	3	1	−2	1	{2, 2, 2, 2}	0	4	1	−3
14	0	0	{1, 1}	2	0	−1	1	0	{2, 2}	0	2	1	−1	1	{2, 2, 2}	0	3	1	−2
15	0	0	{1}	1	0	0	1	0	{2}	0	1	1	0	1	{2, 2}	0	2	1	−1
16	0	0	{ }	0	0	1	1	0	{ }	0	0	1	1	1	{2}	0	1	1	0
17	0	1	{ }	0	0	1	1	1	{ }	0	0	1	1	1	{ }	0	0	1	1
18	0	2	{ }	0	0	1	1	2	{ }	0	0	1	1	2	{ }	0	0	1	1
19	0	3	{ }	0	0	1	1	3	{ }	0	0	1	1	3	{ }	0	0	1	1

facility operating from a warehouse and supplying a single product to m DCs. Without loss of generality, we prioritize the DCs assuming that DC 1 has the highest priority and DC m has the lowest priority. We assume that no prioritization is considered at DCs.

Table 1 shows the inventory and backlog levels of a system with a centralized warehouse operating from a manufacturer and two DCs, where the transportation times between the warehouse and the DCs are zero. In this table, column E represents the events occurring at the warehouse, where 0 indicates production, 1 and 2 indicate an order arrival from DC 1 and DC 2, respectively; column I_0 is the inventory level at the warehouse, which is nonnegative; column B is the set of backlogged orders at the warehouse; columns B_1 and B_2 are the number of backlogs for DC 1 and DC 2 at the warehouse, respectively; and I_1 and I_2 are the inventory level at DC 1 and DC 2, respectively. To keep track of the number of backlogs at the DCs, we allow negative I_j to represent the number of backlogs at DC j. In contrast, since we capture the backlogs of orders at the warehouse in B, we have $0 \le I_0 \le S_0$. The base-stock levels under all policies are $S_0 = 3$, $S_1 = S_2 = 1$, starting with $I_0 = 3$, $I_1 = I_2 = 1$. Note that these values and those for the rationing levels, were chosen for ease of exposition (not as optimal controls of each policy).

3.1. FCFS Policy

Based on the *first-come, first-served* policy, orders at the warehouse are served in order of their arrival. Therefore, completed items are allocated to the DC whose order has waited the longest time in the system.

For example, in Table 1 (column B), under the FCFS policy, the first backlog at the warehouse at event 5 is for DC 2. This is the first backlog to be satisfied at event 11. Note that at event 14, a backlog for DC 2 is satisfied. This increases the inventory level at DC 2 even though there are backlogged customers at DC 1. Such allocation is clearly suboptimal when transportation time is zero (as in this example); nevertheless, this is the FCFS allocation policy.

3.2. SP Policy

Under the *strict priority* policy, orders at the warehouse are satisfied on a FCFS basis as long as the stock level at the warehouse is positive. Otherwise, when there are backlogs at the warehouse, orders are prioritized such that DC 1's orders have the highest priority and DC m's orders have the lowest priority.

For example, at event 11 in Table 1, under the SP policy, the order from DC 1 that has waited (among DC 1's backlogs) the longest time at the warehouse, i.e., arrived at event 6, is satisfied. Note that under the SP policy, although there is a backlog for DC 2 at the warehouse that occurred at event 5, backlogs for DC 1 at the warehouse are satisfied first. Note also that in this example, the SP policy differs from the FCFS policy only starting at event 11, i.e., until the first allocation during a period with backlogs.

3.3. MR Policy

Under the *multilevel rationing* policy, the warehouse has nonnegative and nondecreasing threshold levels R_r, $r = 1, \ldots, m+1$ with $R_1 = 0$, $R_2 \ge 0$, and $R_{m+1} = S_0$, the base-stock level at the warehouse. If the inventory level, I_0, is between R_{r+1} and $R_r + 1$, i.e., $R_r < I_0 \le R_{r+1}$, only orders from DC 1 to DC r are satisfied from the warehouse on a FCFS basis, and orders from the

other DCs are backlogged at the warehouse. When the inventory level is below R_2, only orders from DC 1 are satisfied. When a production ends, backlogs for DCs $r = 1, \ldots, m$ are served if $I_0 = R_r$. If $R_r = R_{r+1}$, backlogs for DC r are satisfied before backlogs for DC $r+1$.

Note that de Véricourt et al. (2002) show that the MR policy is the optimal policy in centralized $M/M/1$ make-to-stock systems. Also note that the SP policy is a special case of the MR policy where all the rationing levels except the last one are 0, i.e., $R_1 = R_2 = \cdots = R_m = 0$. ABB show that the MR policy results in a lower cost than the SP and FCFS policies in centralized $M/G/1$ make-to-stock systems.

In the example presented in Table 1, $R_1 = 0$ as required by the MR policy, $R_2 = 1$, and $R_3 = S_0 = 3$. When $0 < I_0 \leq 1$, only DC 1's orders are satisfied from products available at the warehouse. Therefore, at events 4 and 5, although the inventory level at the warehouse is positive, i.e., $I_0 = 1$, orders from DC 2 are backlogged at the warehouse. Moreover, at event 13, the finished product is kept at the warehouse and the inventory level increases even though there are backlogs for DC 2 at the warehouse. Next, after inventory level increases to $R_2 = 1$, the backlogs of DC 2 are satisfied (events 14–17) and only then the inventory level increases from $R_2 = 1$ to $R_3 = S_0 = 3$ (events 18 and 19). In this example, the MR policy differs from the FCFS (and SP) policy starting at event 4, when an item is reserved for a high-priority arrival and is not allocated to a low-priority one.

3.4. Generalized Multilevel Rationing Policy

Because DCs hold inventory, a backlog at the warehouse does not necessarily reflect backlogs at the DCs. Therefore, when there is a backlog at the warehouse and the SP and MR policies prioritize the high-priority DC, this DC may still have inventory, whereas the low-priority DC faces real backlogs. Clearly, such allocation is not optimal. In our MR example in Table 1, at event 13, the inventory level at the warehouse increases to 1 (reserving a product for a future order from DC 1) although there are three backlogs at DC 2 and the inventory level at DC 1 is 1. Our numerical examples in §6 show that the FCFS policy may result in a lower cost than SP and MR policies.

We therefore introduce a new priority policy called generalized multilevel rationing. The idea behind the GMR policy is to differ the prioritization of the DCs. Specifically, the GMR policy allows a DC to be prioritized when the inventory level hits a negative threshold, i.e., when the number of backlogs at the warehouse hits a threshold. Therefore, implementing the GMR policy is as simple as implementing the MR policy.

Under the GMR policy, the warehouse has nondecreasing threshold levels R_r, $r = 1, \ldots, m+1$ with $R_1 = -\infty$, $R_2, \ldots, R_m \in (-\infty, R_{m+1}]$ and $R_{m+1} = S_0$, the base-stock level at the warehouse. In the GMR policy, the rationing levels are allowed to be *negative*. For simplicity, we present the GMR policy for a system with two DCs and explain this policy for the example in Table 1. We describe and analyze the GMR policy for a system with m DCs in Abouee-Mehrizi et al. (2013).

An important feature of the GMR policy is that the FCFS, SP, and MR policies are special cases of the GMR policy: the FCFS with $R_2 = \cdots = R_m = -\infty$, the SP with $R_2, \ldots, R_m = 0$, and the MR with $R_2, \ldots, R_m \geq 0$. Therefore, the optimal cost under the GMR policy is lower than under any of these three policies.

For a system with two DCs, in the GMR policy, there are three rationing levels: $R_1 = -\infty$, $R_2 \in Z$, which may be negative, and $R_3 = S_0 \geq 0$, the base-stock level at the warehouse. If $R_2 \geq 0$, the GMR policy is identical to the MR policy, thus we assume $R_2 < 0$. If $R_2 < 0$, under the GMR policy the first $-R_2$ backlogs at the warehouse are served based on the FCFS policy. Let B_{2^+} denote the number of these backlogs. As soon as the number of backlogs at the warehouse increases to $-R_2$, i.e., $B_{2^+} = -R_2$, the allocation policy changes. Let $\bar{B}_j$ denote the number of backlogs of DC $j = 1, 2$ that find at least $-R_2$ backlogs at the warehouse upon their arrival and are thus prioritized. As long as $\bar{B}_1 > 0$, priority is given to orders from DC 1 over orders from DC 2, if $\bar{B}_1 = 0$ and $\bar{B}_2 > 0$ priority is given to orders from DC 2. With this prioritization, orders from DC 2 that arrive when the number of backlogs is $-R_2$ or more are given higher priority than orders of DC 1 that arrive when the number of backlogs is less than $-R_2$.

Note that a generalized SP policy would prioritize when the number of backlogs exceeds a threshold rather than prioritizing when backlogs exist. When there are two DCs, the generalized SP policy is identical to the GMR policy. When there are more than two DCs, the generalized SP policy is a special case of the GMR policy with identical $R_2, \ldots, R_m$, i.e., $R_1 = -\infty$, $R_2 = R_3 = \cdots = R_m \in (-\infty, R_{m+1}]$ and $R_{m+1} = S_0$. We thus do not discuss the generalized SP policy any further.

As we show in the example below and in §4, the analysis of the GMR policy is more involved than that of the MR policy, despite the simple difference in their parameters. The main difficulty is to keep track of the backlogs from different DCs at the warehouse.

The GMR allocation policy for the sequence of events in Table 1 is demonstrated in Table 2. Column B_{2^+} is the set of backlogs in Table 2, and columns B_1 and B_2 denote the number of backlogs of DC 1 and DC 2. In this table, $R_2 = -3$, $R_3 = S_0 = 3$. Therefore, the first three backlogs at the warehouse (events 5–7) are included in B_{2^+}. Then, backlogged orders increase $\bar{B}_j$, (events 8–10). Note that until event 10, the GMR policy is identical to both the FCFS and SP polices. They differ once production ends and there are more than $-R_2$ $(= 3)$ backlogs.

Table 2 Stock Allocation Under the LQF and GMR Policies

		GMR									LQF					
Event	E	I_0	B	B_{2+}	$\bar{B}_2$	$\bar{B}_1$	B_1	B_2	I_1	I_2	I_0	B	B_1	B_2	I_1	I_2
1		3	{}	0	0	0	0	0	1	1	3	{}	0	0	1	1
2	1	2	{}	0	0	0	0	0	1	1	2	{}	0	0	1	1
3	1	1	{}	0	0	0	0	0	1	1	1	{}	0	0	1	1
4	2	0	{}	0	0	0	0	0	1	1	0	{}	0	0	1	1
5	2	0	{2}	{2}	0	0	0	1	1	0	0	{2}	0	1	1	0
6	1	0	{2, 1}	{2, 1}	0	0	1	1	0	0	0	{2, 1}	1	1	0	0
7	2	0	{2, 1, 2}	{2, 1, 2}	0	0	1	2	0	−1	0	{2, 1, 2}	1	2	0	−1
8	2	0	{2, 1, 2, 2}	{2, 1, 2}	1	0	1	3	0	−2	0	{2, 1, 2, 2}	1	3	0	−2
9	1	0	{2, 1, 2, 2, 1}	{2, 1, 2}	1	1	2	3	−1	−2	0	{2, 1, 2, 2, 1}	2	3	−1	−2
10	1	0	{2, 1, 2, 2, 1, 1}	{2, 1, 2}	1	2	3	3	−2	−2	0	{2, 1, 2, 2, 1, 1}	3	3	−2	−2
11	0	0	{2, 1, 2, 2, 1}	{2, 1, 2}	1	1	2	3	−1	−2	0	{2, 2, 2, 1, 1}	2	3	−1	−2
12	0	0	{2, 1, 2, 2}	{2, 1, 2}	1	0	1	3	0	−2	0	{2, 2, 1, 1}	2	2	−1	−1
13	0	0	{2, 1, 2}	{2, 1, 2}	0	0	1	2	0	−1	0	{2, 2, 1}	1	2	0	−1
14	0	0	{1, 2}	{1, 2}	0	0	1	1	0	0	0	{2, 1}	1	1	0	0
15	0	0	{2}	{2}	0	0	0	1	1	0	0	{2}	0	1	1	0
16	0	0	{}	0	0	0	0	0	1	1	0	{}	0	0	1	1
17	0	1	{}	0	0	0	0	0	1	1	1	{}	0	0	1	1
18	0	2	{}	0	0	0	0	0	1	1	2	{}	0	0	1	1
19	0	3	{}	0	0	0	0	0	1	1	3	{}	0	0	1	1

When a production ends, backlogs of type $\bar{B}_1$ are served first (events 11, 12). Once $\bar{B}_1 = 0$, backlogs of type $\bar{B}_2$ are also satisfied (event 13) even though there is a backlog of DC 1 at the warehouse (in B_{2+}). The backlogs in B_{2+} are then satisfied and inventory is raised to the base-stock level in events 14–16 and 17–19, respectively.

Note that in events 12–14 of this example, the backlog for DC 1 at the warehouse does not represent a backlogged customer at this DC. Therefore, the GMR policy does not increase the inventory level at DC 1 by much when there are many backlogged customers at DC 2.

Note also that whereas the order filled at event 11 of the SP policy is the first type 1 order backlogged at the warehouse, the order filled at this event under the GMR policy is the first type 1 order arriving after the number of backlogs at the warehouse was 3 ($= -R_2$) or more, i.e., the second overall type 1 order to be backlogged at the warehouse. Therefore, the GMR policy does not necessarily fulfill orders of the same type at the warehouse in a FCFS basis. However, since there is a single product and items are only allocated to customers at the DCs, the DCs still satisfy customers based on the FCFS policy. Therefore, this assumption of the GMR policy does not violate the FCFS allocation to customers at the DCs, which helps to keep the analysis tractable.

3.5. LQF Policy

Under the *longest queue first* policy, orders at the warehouse are satisfied on a FCFS basis as long as the stock level at the warehouse is positive. Otherwise, when there are backlogs at the warehouse, orders are prioritized such that orders from the DC with the highest number of backlogs are served first. Thus, this policy ignores the difference in backlog and holding costs among the DCs. When the number of backlogs are equal, the priority is assigned randomly. (This randomization leads to lower costs in most of our numerical examples.)

The LQF policy is shown in Table 2. For example, at event 12 in Table 2, under the LQF policy, the order from DC 2 that has a higher number of backlogs at the warehouse is satisfied.

3.6. Myopic (T)

Myopic policies, unlike the other policies discussed so far, use information about the current inventory levels at DCs to determine the high-priority DC at any given time. These policies help investigate the effect of the inventory information at the DCs on the total cost of the system. There are several versions of myopic policies. Pena Prez and Zipkin (1997) compare the myopic (P) policy with the myopic (T) policy and show numerically that the latter is often better. The difference between these policies is the length of their look-ahead horizon. We consider the myopic (T) with a look-ahead horizon as the steady-state flow time of units to each DC. This flow time is defined as the steady-state time elapsed from the time an order is placed at a DC until the product corresponding to this order is delivered, assuming all future allocations are made to this DC. We denote the distribution of the number of arrivals during the flow time of a unit to DC j by $F_j(\cdot)$. In §6.1, we derive $F_j(\cdot)$ for all j using the QD approach while assuming there is no order crossing in time.

Given $F_j(\cdot)$, the myopic (T) policy operates as follows: when there is inventory at the warehouse, orders are satisfied in a FCFS fashion. Otherwise, orders are prioritized. At the time of production completion, the inventory position (current inventory plus inventory in transit) at each DC j, IO_j, is observed. Then, the order is sent to the DC with the lowest value of

$$-c_j(1-F_j(IO_j))+h_jF_j(IO_j). \quad (2)$$

3.7. Specifying Priorities

The policies that we consider focus on prioritizing according to the backlog costs c_j. Specifically, we prioritize DC 1 if $c_1 > c_2$. However, in some situations prioritizing based on other parameters, such as the holding costs h_j or the ratio of the holding to backlog costs h_j/c_j, may be useful. We analyze the system for given priorities of the DCs that can be based on any parameter. Therefore, our results can be used to characterize the exact costs under the SP, MR, and GMR policies where priorities are dictated by other parameters.

An example is the *MRh* policy. This is an MR policy where DCs are prioritized based on their holding costs. If the holding costs of two DCs are equal, they are prioritized based on their h_j/c_j ratios. The MRh policy works well when h_1 is small and h_2 is large. In most of the numerical examples considered in §6, this policy does not perform well. Consequently, in the rest of the paper we focus on priority policies based on the backlog cost c_j.

4. Distribution Centers' Cost Functions and Optimal Base-Stock Levels

In this section we analyze the cost of DC j, $C_j^\bullet(S_0, S_j)$, for a given base-stock level at the warehouse, S_0, and at DC j, S_j, under the different allocation policies at the warehouse. We devote §5 to the relevant costs at the warehouse. We use the QD approach to derive the total cost of the DCs and characterize their optimal base-stock levels.

For given S_0, let $I_j^\bullet$ denote the steady-state inventory level at DC j, where $I_j^\bullet < 0$ denotes backlogs at DC j. (For convenience we omit the dependency of I_0 and other random variables defined here on S_0 from the notation.) To express $I_j^\bullet$, we define the shortfall process $N_j^\bullet$ as,

$$N_j^\bullet = S_j - I_j^\bullet. \quad (3)$$

Note that the shortfall process describes the number of outstanding orders at DC j and is a standard method to analyze a single class make-to-stock queue with base-stock level control (see, e.g., Baron 2008, and references therein). For a given S_j, if we know the steady-state distribution of $N_j^\bullet$, we can obtain the steady-state distribution of $I_j^\bullet$ using (3) and calculate the total cost of a DC using Theorem 1 below. Let $P_j^\bullet(i)$ and $E(N_j^\bullet)$ denote the steady-state probability of having i outstanding orders and the expected number of outstanding orders at DC j, respectively. Then, we have the following:

Theorem 1. *Given the base-stock levels S_0 and S_j, the long-run average cost of DC j is*

$$C_j^\bullet(S_0,S_j)=(h_j+c_j)\sum_{x=0}^{S_j-1}(S_j-x)P_j^\bullet(x)+c_jE(N_j^\bullet)-c_jS_j, \quad (4)$$

and the optimal base-stock level at DC j, $(S_j^\bullet)^*$, is

$$(S_j^\bullet)^* = \min\left\{k:\ \sum_{r=0}^{k}P_j^\bullet(r) > c_j/(h_j+c_j)\right\}. \quad (5)$$

Although Theorem 1 provides the optimal S_j given $P_j^\bullet(x)$ for each $x \geq 0$, the heart of the matter is, of course, expressing these probabilities. We therefore investigate $P_j^\bullet(x)$, the steady-state distribution of $N_j^\bullet$. This will also allow us to derive closed-form expressions for $E(N_j^\bullet)$, considering that $E(N_j^\bullet) = -(d\Pi_j^\bullet(z)/dz)|_{z=1}$, where $\Pi_j^\bullet(z)$ denotes the steady-state probability generating function (PGF) of $N_j^\bullet$.

Let $N_{L_j}^\bullet$ and $N_{W_j}^\bullet$ denote the steady-state number of outstanding orders for DC j that are in transit and at the warehouse, respectively. And, let $\Pi_{L_j}^\bullet(z)$ and $\Pi_{W_j}^\bullet(z)$ denote the PGF of the distribution of $N_{L_j}^\bullet$ and $N_{W_j}^\bullet$, respectively. Svoronos and Zipkin (1991) show that $N_j^\bullet$ is the convolution of the steady-state distributions of $N_{L_j}^\bullet$ and $N_{W_j}^\bullet$, and under the FCFS policy, the PGF of $N_j^\bullet$ can be obtained using

$$\Pi_j^\bullet(z) = \Pi_{L_j}^\bullet(z)\Pi_{W_j}^\bullet(z). \quad (6)$$

In Lemmas 1 and 2 in §4.1 and 4.2, respectively, we show that the relation (6) holds under the SP, MR, and GMR polices. As a result, the PGF of the number of outstanding orders in transit is

$$\Pi_{L_j}^\bullet(z) = \tilde{L}_j(\lambda_j(1-z)). \quad (7)$$

In the next section, we focus on deriving the PGF of the number of DC j's orders at the warehouse, $\Pi_{W_j}^\bullet(z)$, under the different policies. The derivation of $\Pi_{W_j}^\bullet(z)$ for the FCFS, SP, and MR policies in §4.1 follows the derivation in ABB. In §4.2, we extend the methodology from ABB to the GMR policy.

4.1. FCFS, SP, and MR Policies

In this section, we investigate the steady-state distribution of the number of outstanding orders at DC j under the FCFS, SP, and MR policies. We first demonstrate that relation (6) holds under these allocation policies.

Lemma 1. *The PGF of the number of outstanding orders from DC j under the FCFS, SP, and MR policies can be obtained using*

$$\Pi_j^\bullet(z) = \Pi_{L_j}^\bullet(z)\Pi_{W_j}^\bullet(z). \quad (8)$$

To obtain the PGF of the distribution of $N^{\bullet}_{W_j}$, we first characterize the steady-state LT of the waiting time distribution of a DC j order at the warehouse, $\tilde{w}^{\bullet}_j(s)$. Then, we obtain the PGF of the distribution of $N^{\bullet}_{W_j}$ using Little's distributional law, as in Bertsimas and Nakazato (1995):

$$\Pi^{\bullet}_{W_j}(z) = \tilde{w}^{\bullet}_j(\lambda_j(1-z)). \tag{9}$$

To obtain $\tilde{w}^{\bullet}_j(s)$, we consider two cases: (1) an order from DC j arriving at the warehouse is satisfied immediately, and (2) an arriving order is backlogged.

Case 1. Recall that when a customer arrives at DC j, a new unit of the product is immediately ordered from the warehouse; it is satisfied immediately if the inventory level at the warehouse is greater than R_j for $j = 2, \ldots, m$ under the FCFS, SP, and MR policies. Therefore, $\tilde{w}^{\bullet}_j(s \mid I_0 > R_j)$ is equal to 1, i.e., the waiting time at the warehouse of an order from DC j is zero if the inventory level at the warehouse is greater than R_j upon the order's arrival.

Case 2. Consider the case where upon the arrival of a demand from DC j, the warehouse is out of stock. Since $\tilde{w}^{\bullet}_j(s \mid I_0 \le R_j)$, the LT of the steady-state waiting time of orders of DC j at the warehouse, depends on the allocation policy; thus, we study it for each allocation policy separately.

Let $P^{\bullet}(I_0 > R_j)$ denote the probability that an order from DC j to the warehouse is satisfied immediately. Then, the LT of the steady-state waiting time distribution of an order of DC j, $\tilde{w}^{\bullet}_j(\cdot)$ is

$$\tilde{w}^{\bullet}_j(s) = P^{\bullet}(I_0 > R_j) + (1 - P^{\bullet}(I_0 > R_j))\tilde{w}^{\bullet}_j(s \mid I_0 \le R_j). \tag{10}$$

We can thus obtain the PGF of the distribution of the total number of outstanding orders at DC j, $\Pi^{\bullet}_j(z)$ by substituting (10) into (6):

$$\begin{aligned}\Pi^{\bullet}_j(z) = {} & P^{\bullet}(I_0 > R_j)\tilde{L}_j(\lambda_j(1-z)) + (1 - P^{\bullet}(I_0 > R_j)) \\ & \cdot \tilde{w}^{\bullet}_j(\lambda_j(1-z) \mid I_0 \le R_j)\tilde{L}_j(\lambda_j(1-z)).\end{aligned} \tag{11}$$

Note that $\Pi^{\bullet}_j(z)$ can be numerically inverted to obtain $P^{\bullet}_j(i)$ (see, e.g., Abate and Whitt 1992).

Given the rationing levels R_j for $j = 2, \ldots, m$, to obtain all elements in (11), we only need to characterize the probability of backlog, $P^{\bullet}(I_0 \le R_j)$, and $\tilde{w}^{\bullet}_j(s \mid I_0 \le R_j)$ under the FCFS, SP, and MR policies. These derivations are provided next.

4.1.1. FCFS Policy. To obtain $\tilde{w}^{\mathrm{FCFS}}_j(s \mid I_0 \le R_j) = \tilde{w}^{\mathrm{FCFS}}_j(s \mid I_0 = 0)$ under the FCFS policy, we define the FCFS backlog queue to include periods of time when there is no inventory in the system. This queue is different from an $M/G/1$ queue since when the inventory level decreases to zero and a new period with no inventory begins, there are S_0 outstanding orders at the production facility; hence the server is busy. This FCFS backlog queue is similar to the SP backlog queue discussed in Online Appendix B, but here the allocation policy is FCFS; see §3.2.1 in ABB for further discussion.

Using the FCFS backlog queue, we can characterize $\tilde{w}^{\mathrm{FCFS}}_j(s \mid I_0 = 0)$. Let ρ_b be the server utilization in the FCFS backlog queue, and $1/\mu_1$ be the first moment of the exceptional first service times; $1/\mu_1$ can be obtained using $\tilde{b}_{S_0}(s)$ in this queue ($\tilde{b}_j(s)$ is given in (B.37) in Online Appendix B). Then, ρ_b can be obtained from (B.38) in Online Appendix B.

Theorem 2. *The LT of the steady-state waiting time distribution of an order from DC j at the warehouse that finds the warehouse out of stock under the FCFS allocation policy is*

$$\tilde{w}^{\mathrm{FCFS}}_j(s \mid I_0 = 0) = \frac{(1-\rho_b)(\lambda(\tilde{b}(s) - \tilde{b}_{S_0}(s)) + s\tilde{b}_{S_0}(s))}{s - \lambda(1 - \tilde{b}(s))}. \tag{12}$$

In steady-state, an order from DC j that finds the warehouse out of stock upon its arrival encounters a delay with a LT of $\tilde{w}^{\mathrm{FCFS}}_j(s \mid I_0 = 0)$. After this delay, the order is satisfied and spends τ_j units of time in transit with a LT of $\tilde{L}_j(s)$ before arriving at the DC.

Since DC j places a new order to the warehouse as soon as a demand arrives to the DC, the probability that an order from DC j finds the warehouse empty is $P^{\mathrm{FCFS}}(I_0 \le R_j)$. Because $R_j = 0$ under the FCFS allocation policy, this probability is identical to the probability that the number of orders in a single class $M/G/1$ make-to-stock system with the FCFS policy is greater than $S_0 - 1$ denoted by $\bar{F}^{\mathrm{FCFS}}(S_0 - 1)$. Letting $\rho = \lambda/\mu$, then the probability of having i people in a single class $M/G/1$ make-to-stock system with the FCFS policy is (see Equation (4) in ABB)

$$(1-\rho)\prod_{j=0}^{i-1}\frac{1-\tilde{b}_j(\lambda)}{\tilde{b}(\lambda)},$$

and the probability that the number of orders in this system is greater than $S_0 - 1$ is

$$\bar{F}^{\mathrm{FCFS}}(S_0 - 1) = 1 - \sum_{i=0}^{S_0-1}(1-\rho)\prod_{j=0}^{i-1}\frac{1-\tilde{b}_j(\lambda)}{\tilde{b}(\lambda)}. \tag{13}$$

Substituting (12) and (13) into (11), we get the PGF of the steady-state distribution of outstanding orders at DC j, $\Pi^{\mathrm{FCFS}}_j(z)$. We can then obtain the optimal cost and base-stock level at a DC for a given base-stock level at the warehouse using Theorem 1. Furthermore, using (11) and that $E(N^{\mathrm{FCFS}}_j) = -(d\Pi^{\mathrm{FCFS}}_j(z)/dz)|_{z=1}$, for the FCFS policy (after some algebra), we get

$$\begin{aligned}E(N^{\mathrm{FCFS}}_j) = {} & \frac{\lambda_j}{\theta_j} + \bar{F}^{\mathrm{FCFS}}(S_0 - 1)(1-\rho_b)\lambda_j \\ & \cdot \frac{(\lambda)^2 m_2/\mu_1 + (1-\rho)(\lambda m_2^1 + 2/\mu_1)}{2(1-\rho)^2},\end{aligned} \tag{14}$$

which can be substituted into (4), where m_2 and m_2^1 are the second moments of $\tilde{b}(\cdot)$ and $\tilde{b}_{S_0}(\cdot)$, respectively.

4.1.2. SP Policy. To calculate $\tilde{w}_j^{\mathrm{SP}}(s \mid I_0 = 0)$, we note that this LT is identical to the LT of the steady-state waiting time distribution of class j customers in an $M/G/1$ make-to-stock queue with prioritization. The LT of this distribution is given in (B.39) in Online Appendix B.

Let $\bar{F}^{\mathrm{SP}}(S_0 - 1) = P^{\mathrm{SP}}(I_0 = 0)$ denote the probability that an order from DC j finds the warehouse out of stock under the SP policy. Since the orders are satisfied in the order of their arrival under both SP and FCFS policies as long as the inventory level at the warehouse is positive, the probability that an order from DC j finds the warehouse out of stock under the SP policy is identical to the one under the FCFS policy, i.e., $\bar{F}^{\mathrm{SP}}(S_0 - 1) = \bar{F}^{\mathrm{FCFS}}(S_0 - 1)$, as given in (13).

Substituting $\tilde{w}_j^{\mathrm{SP}}(s \mid I_0 = 0)$ given in (B.39) and (13) in (11), we obtain $\Pi_j^{\mathrm{SP}}(z)$, the PGF of the steady-state distribution of outstanding orders at DC j, so that we can express the optimal cost and base-stock level at a DC for a given base-stock level at the warehouse under the SP policy. Using $E(N_j^{\mathrm{SP}}) = -(d\Pi_j^{\mathrm{SP}}(z)/dz)|_{z=1}$, we can substitute

$$E(N_j^{\mathrm{SP}}) = \frac{\lambda_j}{\theta_j} + \bar{F}^{\mathrm{SP}}(S_0 - 1)(1 - \rho_b)\frac{\lambda_j(1-\rho)}{(1-\lambda_j^+/\mu)(1-\lambda_{j-1}^+/\mu)} \cdot \frac{\lambda^2 m_2/\mu_1 + (1-\rho)(\lambda m_2^1 + 2/\mu_1)}{2(1-\rho)^2} \tag{15}$$

into (4), where m_2 and m_2^1 are the second moments of $\tilde{b}(\cdot)$ and $\tilde{b}_{S_0}(\cdot)$, respectively.

4.1.3. MR Policy. In this section, we determine $\tilde{w}_j^{\mathrm{MR}}(s)$. Note that under the MR policy, unlike the FCFS and SP policies, an order arriving at the warehouse from DC j may be backlogged even if the inventory level is positive. In general, an order from DC j is served from on-hand inventory if the inventory level at the warehouse is above R_j.

ABB analyze a single product multiclass $M/G/1$ make-to-stock queue with the MR policy and characterize the LT of the steady-state waiting time distribution of a class j arrival that finds the inventory level less than $R_j + 1$ as given in (B.41) in Online Appendix B. To use the results in ABB (given in Online Appendix B), we define the jth backlog queue, BQ_j, as a two-priority queue with an exceptional first service time in each busy period and a utilization of ρ_b^j, where orders of DCs $1, \ldots, j-1$ are high priority and orders from DC j are low priority. This queue corresponds to the original system during periods when orders from DC j are backlogged at the warehouse. We denote the LT of the exceptional first service times in the jth backlog queue, BQ_j, by $\tilde{b}_{\Delta_j}^j(s)$, where $\Delta_j = R_{j+1} - R_j$. This LT can be obtained using Algorithm 1 given in ABB. Then, $\tilde{w}_j^{\mathrm{MR}}(s \mid I_0 \le R_j)$ can be obtained using (B.41).

Substituting (B.41) and (B.44) of Online Appendix B in (11), we express $\Pi_j^{\mathrm{MR}}(z)$, the PGF of the steady-state distribution of outstanding orders at DC j, and use it to obtain the optimal cost and base-stock level at a DC for given rationing levels at the warehouse under the MR allocation policy. Using $E(N_j^{\mathrm{MR}}) = -(d\Pi_j^{\mathrm{MR}}(z)/dz)|_{z=1}$, we can substitute

$$E(N_j^{\mathrm{MR}}) = \frac{\lambda_j}{\theta_j} + F_j^{\mathrm{MR}}(R_j)(1-\rho_b^j)\frac{\lambda_j(1-\rho)}{(1-\lambda_j^+/\mu)(1-\lambda_{j-1}^+/\mu)} \cdot \frac{(\lambda_j^+)^2 m_2/\mu_1 + (1-\rho)(\lambda_j^+ m_2^1 + 2/\mu_1)}{2(1-\rho)^2}, \tag{16}$$

into (4), where m_2 and m_2^1 are the second moments of $\tilde{b}(\cdot)$ and $\tilde{b}_{\Delta_j}^j(\cdot)$, respectively.

4.2. GMR Policy

In this section, we investigate the steady-state distribution of the number of outstanding orders at DC j under the GMR policy. The analysis of outstanding orders under this policy is more challenging than the previous two priority policies since orders from the low-priority DCs may be served before high-priority ones at the warehouse. To keep the discussion simple, we only consider an inventory system with a warehouse and two DCs and explain how the required probabilities can be obtained. In Abouee-Mehrizi et al. (2013), we generalize the results to a system with $m(>2)$ DCs.

Note that in a system with two DCs, if $R_2 \ge 0$, the GMR policy is identical to the MR policy. Thus, we assume $R_2 < 0$. To characterize the distribution of the number of outstanding orders at DC j, we use the backlog queues defined in §4.1.3. In a system with two DCS, we need to consider the second backlog queue, BQ_2, and the third backlog queue, BQ_3. This third backlog queue, BQ_3, is the standard FCFS single class $M/G/1$ queue with an arrival rate of λ. This queue corresponds to the durations when the number of backlogs is less than $-R_2$, i.e., when the shortfall process is lower than $S_0 + R_2$. The second backlog queue, BQ_2, is a two-class $M/G/1$ priority queue with an exceptional first service time in each busy period where the arrival rates of high- and low-priority orders are λ_1 and λ_2, respectively. This queue corresponds to the durations when the number of backlogs is more than $-R_2 - 1$, i.e., there are backlogs in the system and DCs are prioritized.

Note that under the GMR policy, orders may cross at the warehouse since orders served under the priority policy, $\bar{B}_j$, leave the warehouse before orders served under the FCFS policy, B_2^+. Therefore, the distributional Little's law cannot be applied directly to the system. Still, in each DC (and each backlog queue) backlogs of that queue are served on a FCFS basis. Thus, Little's

distributional law can be used to obtain the required PGF in each backlog queue. In the following lemma, we state this argument precisely.

Lemma 2. *The PGF of the number of outstanding orders at DC j under the GMR policy can be obtained using*

$$\Pi_j^{\mathrm{GMR}}(z) = \Pi_{L_j}^{\mathrm{GMR}}(z)\Pi_{W_j}^{\mathrm{GMR}}(z). \tag{17}$$

Given Lemma 2 and (7), we only need to derive the steady-state distribution of backlogs of DC j at the warehouse, $\Pi_{W_j}^{\mathrm{GMR}}(z)$, to express the distribution of the outstanding orders at the DC. We next obtain this distribution.

Under the GMR policy, as long as the total number of backlogs at the warehouse is not greater than $-R_2$ ($R_2 < 0$), orders are served based on the FCFS policy. Therefore, as in Gayon et al. (2009), the number of type $j = 1, 2$ backlogs is binomially distributed with parameter (λ_j/λ). Thus, when the warehouse is out of stock and the total number of backlogs at the warehouse is $B_{2^+} = i$ ($< -R_2$) we have the following:

$$P(N_{W_j} = n_j \mid B_{2^+} = i, I_0 = 0) = \binom{i}{n_j}\left(\frac{\lambda_j}{\lambda}\right)^{n_j}\left(\frac{\lambda-\lambda_j}{\lambda}\right)^{i-n_j}, \quad i = 0, \ldots, -R_2 - 1,\ n_j = 0, \ldots, i. \tag{18}$$

(Conditioning on $I_0 = 0$ in (18) is only required for $i = 0$.)

Recall that under the GMR policy when the total number of backlogs at the warehouse is greater than $-R_2$, we have $B_{2^+} = -R_2$. The backlogged orders of $\bar{B}_1$ are served first and then orders of $\bar{B}_2$. The backlogs of B_{2^+} will be served in order of their arrival only if the total number of backlogs at the warehouse is not greater than $-R_2$. When there are no backlogs of $\bar{B}_1$ and $\bar{B}_2$ (i.e., the total number of backlogs decreases to $-R_2$), backlogs of B_{2^+} are satisfied. Therefore, when there are more than $-R_2$ backlogs at the warehouse, the policy is similar to the MR and SP policies.

Let $P(B_{2^+}^j = i \mid B_{2^+} = -R_2)$ denote the steady-state probability of having i backlogs from DC j in B_{2^+} given that there are $-R_2$ backlogs. Then, the distribution of the number of backlogs of DC j given that $B_{2^+} = -R_2$ is

$$\begin{aligned} &P(N_{W_j} = n_j \mid B_{2^+} = -R_2) \\ &= \sum_{k=0}^{n_j} P(B_{2^+}^j = k \mid B_{2^+} = -R_2) P(\bar{B}_j = n_j - k \mid B_{2^+} = -R_2) \\ &= \sum_{k=0}^{n_j} \binom{-R_2}{k}\left(\frac{\lambda_j}{\lambda}\right)^{k}\left(\frac{\lambda-\lambda_j}{\lambda}\right)^{-R_2-k} \\ &\quad \cdot P(\bar{B}_j = n_j - k \mid B_{2^+} = -R_2), \quad n_j = 0, 1, \ldots. \end{aligned} \tag{19}$$

We next characterize $P(\bar{B}_j = i)$. Under the GMR policy, orders arriving at the warehouse are served in a FCFS basis as long as the total number of orders is less than $R_3 - R_2$; otherwise they are prioritized. Similarly, under the MR policy orders arriving at the warehouse are prioritized only if the total number of orders at the warehouse is not less than $R_3 - R_2$. Therefore, similar to analysis of the MR policy, we can obtain $P(\bar{B}_j = i)$ using the PGF of $P_j^{BQ_2}(i)$ given in (B.43) in Online Appendix B and Little's distributional law in (9).

Combining (18) and (19), we obtain the following distribution for backlogs of DC j for a system with two DCs if the warehouse is out of stock:

$$\begin{aligned} &P(N_{W_j} = n_j \mid I_0 = 0) \\ &= \sum_{i=0}^{-R_2-1} P(N_{W_j} = n_j \mid B_{2^+} = i, I_0 = 0) P(B_{2^+} = i \mid I_0 = 0) \\ &\quad + P(N_{W_j} = n_j \mid B_{2^+} = -R_2) P(B_{2^+} = -R_2). \end{aligned} \tag{20}$$

Next, we derive $P(B_{2^+} = i)$ for $i < -R_2$, the probability of having $i < -R_2$ backlogs at the warehouse. Let N_0 denote the total number of outstanding orders at the warehouse. Given that $I_0 = 0$ and $B_{2^+} < -R_2$, N_0 can be expressed as

$$N_0 = S_0 + B_{2^+}, \quad I_0 = 0, \quad B_{2^+} < -R_2. \tag{21}$$

But, from the definition of the third backlog queue (the shortfall queue), BQ_3, the total number of outstanding orders at the warehouse, N_0, is identical in distribution to the total number of jobs in BQ_3. Therefore, by conditioning on $I_0 = 0$ (or equivalently $N_0 \geq S_0 = R_3$), we get

$$P(B_{2^+} = i \mid I_0 = 0) = \frac{P^{BQ_3}(i + R_3)}{1 - \sum_{k=0}^{R_3-1} P^{BQ_3}(k)}, \quad i = 0, \ldots, -R_2, \tag{22}$$

where $P^{BQ_3}(i + R_3)$ is the probability that the number of jobs in BQ_3 is equal to $i + R_3$ (BQ_3 is defined in §B.2 of Online Appendix B). Combining (18), (19), and (22), we get the distribution of the number of outstanding orders for DC j at the warehouse for a system with two DCs (for $R_2 < 0$).

Theorem 3. *For the GMR policy, the distribution of the number of outstanding orders for DC j at the warehouse for a system with two DCs and $R_2 < 0$ is*

$$\begin{aligned} &P(N_{W_j} = n_j \mid I_0 = 0) \\ &= \sum_{i=n_j}^{-R_2-1} \binom{i}{n_j}\left(\frac{\lambda_j}{\lambda}\right)^{n_j}\left(\frac{\lambda-\lambda_j}{\lambda}\right)^{i-n_j} \frac{P^{BQ_3}(i + R_3)}{1 - \sum_{y=0}^{R_3-1} P^{BQ_3}(y)} \\ &\quad + \frac{\sum_{i=0}^{-R_2-1} P^{BQ_3}(i + R_3)}{1 - \sum_{y=0}^{R_3-1} P^{BQ_3}(y)} \sum_{i=0}^{n_j} \binom{-R_2}{i}\left(\frac{\lambda_j}{\lambda}\right)^{i} \\ &\quad \cdot \left(\frac{\lambda-\lambda_j}{\lambda}\right)^{R_2-i} P_j^{BQ_2}(n_j - i). \end{aligned} \tag{23}$$

We next characterize the PGF of the distribution of the total number of orders at DC j under the GMR policy, $\Pi_j^{\mathrm{GMR}}(z)$. Note that this distribution is the convolution of the distributions of the number of the products in transit to DC j, and the number of backlogged orders for DC j at the warehouse. Therefore, we have the following:

Corollary 1. *For given rationing levels at the warehouse, the PGF of the number of outstanding orders at DC j under the GMR policy is*

$$\Pi_j^{\mathrm{GMR}}(z) = \tilde{L}_j(\lambda_j(1-z))(1-\bar{F}^{BQ_3}(R_3-1)) + \tilde{L}_j(\lambda_j(1-z)) \cdot \left(\bar{F}^{BQ_3}(R_3-1)\sum_{i=0}^{\infty} P(N_{W_j}=i \mid I_0=0)z^i\right), \quad (24)$$

where $\bar{F}^{BQ_3}(R_3-1) := 1-\sum_{k=0}^{R_3-1} P^{BQ_3}(k)$.

Using (24) and Theorem 1, we can obtain the optimal cost and base-stock level at a DC for given rationing levels at the warehouse under the GMR policy.

5. Cost and Rationing Levels in the Production Facility

In this section, we investigate the total cost of the warehouse under the four allocation policies. Because the backlog costs of the system is captured in the DCs' cost, we only need to characterize the holding cost at the warehouse.

Recall that we model the production facility as a single product multiclass $M/G/1$ make-to-stock queueing system. ABB analyze this system under FCFS, SP, and MR policies. Given the allocation policy and rationing levels, these results provide closed-form expressions for each of these policies. We present these results below. Let $P(i)$ denote the steady-state probability of having i orders in a single class $M/G/1$ queue. Then, for the FCFS and SP policies, the holding cost of the warehouse is given by

$$C_0^{\mathrm{FCFS}}(S) = C_0^{\mathrm{SP}}(S) = h\sum_{x=0}^{S}(S-x)P(x), \quad (25)$$

where using $\tilde{b}_j(\cdot)$ from (B.37),

$$P(i) = (1-\rho)\prod_{j=0}^{i-1}\frac{1-\tilde{b}_j(\lambda)}{\tilde{b}(\lambda)}.$$

Given the rationing levels, the holding cost of the warehouse under the MR policy is (see Theorem 7 in ABB)

$$C_0^{\mathrm{MR}} = h\sum_{r=2}^{m+1}\left[\prod_{j=r+1}^{m+1}\bar{F}_h^{BQ_j}(R_j-R_{j-1}-1) \cdot \sum_{x=0}^{R_r-R_{r-1}-1}(R_r-x)P_h^{BQ_r}(x)\right], \quad (26)$$

where $P_h^{BQ_r}(\cdot)$ can be obtained using (B.43).

We next discuss the total cost of the warehouse under the GMR allocation policy. We derive this cost for a system with m DCs in Abouee-Mehrizi et al. (2013).

Because in a system with two DCs, the GMR policy is identical to the MR policy if $R_2 \geq 0$, the holding cost at the warehouse under the GMR policy is identical to the one under the MR policy and can be obtained using (26) if $R_2 \geq 0$.

Now suppose $R_2 < 0$. In this case, orders from DCs 1 and 2 are served based on the FCFS policy as long as the inventory level at the warehouse is positive. Therefore, the holding cost at the warehouse under the GMR policy is identical to the one under the FCFS and SP policies given in (25) if $R_2 < 0$.

6. Comparison of Policies

In previous sections, we provided the exact solution for computing the optimal cost of the system under the FCFS, SP, MR, and GMR policies assuming no order crossing. In this section, we address three questions: (1) Is the analysis in the previous sections useful in terms of computation times? (2) Does prioritization decrease the total cost of the system, and what is the added value of prioritizing using the GMR policy? (3) How does the GMR with order crossing policy perform when order crossing is allowed? We note that the optimal controls of the GMROC policy may differ from those of the GMR policy (that does not allow order crossing). To answer this question, we compare the GMR, LQF, myopic (T), and GMROC policies with each other and with the optimal policy (when order crossing exists). In answering each of these questions we also investigate the factors that affect the answers. Possible factors are the level of uncertainty in the production and/or transportation times, how busy the production facility is, and the relationships between different cost parameters.

6.1. Set Up of Numerical Study

We consider a system with one manufacturer and two DCs. We assume that the production and transportation times are either deterministic or exponentially distributed. We set the mean production $\mu = 1$, the holding costs at the warehouse and DC 2 $h_0 = h_2 = 0.5$, the backlog cost at DC 1 $c_1 = 10$, and the other parameters are varied as follows: the utilization $\rho = \lambda/\mu \in \{0.5, 0.8, 0.9\}$; the proportion of arrival rates $\lambda_1/\lambda_2 \in \{0.2, 0.5, 1, 2, 5\}$; the proportion of backlog costs $c_1/c_2 \in \{2, 5, 10\}$; the mean of transportation times $1/\theta_1 = 1/\theta_2 \in \{2, 6\}$; and the proportion of holding costs $h_1/h_2 \in \{1, 2\}$. This gives a total of 720 experiments.

Let $(C^{\bullet})^*(\cdot)$ denote the optimal cost of policy $\bullet$. Each of these costs has two parts: holding cost at the warehouse and holding and backlog costs at the DCs. For FCFS, SP, MR, and GMR policies, we calculate the former using the results from §5 and the latter

using the closed-form results from Theorem 1 in §4. For the LQF, myopic (T), and GMROC policies, we obtain both parts using simulation. For the optimal policy we obtained the optimal cost from a dynamic program.

For each case, we calculate the relative gaps between the cost of policy "•" and the cost of the GMR policy:

$$\Delta\bullet := \frac{C^{\bullet} - C^{\mathrm{GMR}}}{C^{\mathrm{GMR}}} \times 100. \tag{27}$$

So a positive $\Delta\bullet$ means that the GMR policy is the better policy, whereas a negative $\Delta\bullet$ means that the "•" policy is the better policy.

We next explain the exhaustive search procedure that we used to obtain the optimal base-stock and rationing levels under different allocation policies. We search for the optimal base-stock and rationing levels at the warehouse, and obtain the optimal base-stock levels at the DCs, S_1^* and S_2^*, using Theorem 1 for any given set of base-stock and rationing levels at the warehouse.

To obtain the optimal base-stock level at the warehouse under the FCFS and SP policies, we vary S_0 from 0 to M such that $M = \min\{i\colon C^{\bullet}(i+2, S_1^*, S_2^*) > C^{\bullet}(i+1, S_1^*, S_2^*) > C^{\bullet}(i, S_1^*, S_2^*)\}$.

Similarly, to obtain the rationing levels under the MR policy, we search over $R_3 = 0 \ldots M$ by varying $R_2 = 0, \ldots, R_3$. For each R_3, we look for the optimal rationing level R_2. We let $M = \min\{i\colon C^{\mathrm{MR}}(R_2^*, i+2, S_1^*, S_2^*) > C^{\mathrm{MR}}(R_2^*, i+1, S_1^*, S_2^*) > C^{\mathrm{MR}}(R_2^*, i, S_1^*, S_2^*)\}$.

To obtain the optimal rationing levels under the GMR policy, we first search $R_2 = 0, -1, \ldots, -N$ and $R_3 = 0, 1, \ldots, M$ such that for the given R_2,

$$\begin{aligned} M = \min\big\{j\colon\ & C^{\mathrm{GMR}}(R_2, j+2, S_1^*, S_2^*) \\ & > C^{\mathrm{GMR}}(R_2, j+1, S_1^*, S_2^*) \\ & > C^{\mathrm{GMR}}(R_2, j, S_1^*, S_2^*)\big\}, \end{aligned}$$

and

$$\begin{aligned} N = \min\big\{i\colon\ & C^{\mathrm{GMR}}(i-2, R_3^*, S_1^*, S_2^*) \\ & > C^{\mathrm{GMR}}(i-1, R_3^*, S_1^*, S_2^*) \\ & > C^{\mathrm{GMR}}(i, R_3^*, S_1^*, S_2^*)\big\}. \end{aligned}$$

Considering that in a system with two DCs the GMR policy is identical to the MR policy when $R_2 \geq 0$, we compare the minimum total cost found by this search with the optimal cost calculated for the MR policy and choose the one with lower cost as the solution of the GMR policy.

Note that the above procedure does not guarantee obtaining the optimal controls and cost of the system because of the termination condition (it stops whenever it finds two consecutive increases in cost), the sequential procedure used to obtain the rationing levels, and the use of numerical inversion methods to determine required probabilities. Nevertheless, we obtain the best possible solution under the different allocation policies in a reasonable amount of time.

To obtain the optimal base-stock levels for the LQF, myopic (T), and GMROC policies, we use simulation. The cost is calculated from the simulation based on the allocation rule of each policy. For the myopic (T), the allocation rule requires calculating $F_j(\cdot)$, the distribution of the number of arrivals to DC j during a unit flow time. We calculate this distribution assuming that there is no order crossing using (6) with $w_j(\cdot)$ as given in (12). Note that although we assume no order crossing to obtain $F_j(\cdot)$ and make an allocation decision, allocated orders may cross during the transportation time in the simulation. Therefore, the base-stock levels and cost obtained for the myopic (T) policy using simulation is for a system with order crossing. Note also that without the QD approach we could not express $w_j(\cdot)$.

Finally, to obtain the optimal policy, we model the system using a dynamic program and apply the value iteration algorithm (Sennott 1999) to calculate the optimal cost when both production and transportation times are exponential (and therefore order crossing may occur).

6.2. Comparison of Computational Times

The detailed results for the 720 cases show that the computational time required to obtain the optimal cost under the FCFS policy varies from 0.11 to 220.92 seconds with an average of 23.09 seconds; under the SP policy, it varies from 0.2 to 191.59 seconds with an average of 17.49 seconds; and under the GMR policy, it varies from 0.33 to 31,921.92 seconds (8.52 hours) with an average of 326.40 seconds. We find these average times acceptable.

We obtain the optimal costs under the LQF, myopic (T), and GMROC policies using simulation. Our rudimentary simulation (i.e., we do not use any optimization feature of the software to find the optimal controls) sometimes took a very long time (more than 10 hours) to find the optimal control and none of the simulation runs took less then two hours. In view of these long simulation times, and as detailed in the next subsections, we only consider 48 cases for these policies. Obtaining the cost of the optimal policy using a dynamic program also took a very long time. For example, for the cases with $\rho \geq 0.7$, the value iteration algorithm could not find the optimal cost within 72 hours. To formulate the problem as a dynamic program and obtain the optimal policy, we keep track of the inventory level at the warehouse, the number of backlogs of each DC at the warehouse, and the inventory on transit to each DC. Truncating inventory and backlogs at 20, we have $20^5 = 3{,}200{,}000$ possible states. In view of long

run times, we only consider six cases for the optimal policy.

We observe that computational times increase with utilization. This is because higher utilization implies higher base-stock levels, increasing the range of the exhaustive search.

The results in this subsection indicate that our exact derivations are practical in terms of computational times for almost all combinations of parameters, but the computational times of finding the optimal LQF and myopic (T) costs as well as the optimal policy are not comparable with the times required for the other policies.

6.3. Importance of Prioritization

In this section, we discuss whether prioritization is helpful by comparing the performance of the FCFS, SP, MR, and GMR policies first analytically and then numerically. These results are based on our analytical derivation with no order crossing.

6.3.1. Theoretical Comparison. Recall from §3 that the SP policy is a special case of the MR policy and that both the MR and FCFS policies are special cases of the GMR policy. Therefore, we have the following:

Observation 1. In a two-echelon inventory system, we have $C^{\mathrm{GMR}^*}(\cdot) \le C^{\mathrm{MR}^*}(\cdot) \le C^{\mathrm{SP}^*}(\cdot)$, and $C^{\mathrm{GMR}^*}(\cdot) \le C^{\mathrm{FCFS}^*}(\cdot)$.

6.3.2. Numerical Comparison. The summary of the comparisons between the performance of the FCFS, SP, MR, and GMR polices are given in Tables 3 and 4. The first row of each table indicates the distributions of (production, transportation) times. These tables show how the average relative gaps change when a parameter changes. Over the 720 cases, the average cost savings of the GMR with respect to the FCFS is 9.08% with a minimum and a maximum of 0.01% and 64.02%, respectively (the detailed numerical examples are not provided here). This demonstrates the value of prioritizing using the GMR policy.

Table 3 **Average Relative Gaps When Transportation Times Are Deterministic**

		(Det., Det.)			(Exp., Det.)		
		$\Delta FCFS$	ΔSP	ΔMR	$\Delta FCFS$	ΔSP	ΔMR
h_1	0.5	8.44	1.32	1.32	12.63	0.97	0.97
	1	9.62	0.29	0.29	15.19	0.78	0.51
$1/\theta_i$	2	11.04	0.61	0.61	15.57	0.61	0.40
	6	7.02	0.99	0.99	12.25	1.13	1.07
c_1/c_2	2	2.99	1.65	1.65	4.18	1.24	1.24
	5	8.92	0.59	0.59	13.99	0.90	0.82
	10	15.18	0.16	0.16	23.56	0.48	0.15
ρ	0.5	1.66	0.23	0.23	2.83	0.41	0.41
	0.8	7.59	1.01	1.01	13.39	0.88	0.74
	0.9	17.84	1.16	1.16	25.51	1.34	1.06
λ_1/λ_2	0.2	4.51	0.38	0.38	7.17	0.73	0.73
	0.5	7.15	0.71	0.71	12.62	1.09	1.09
	1	9.68	0.97	0.97	16.04	0.74	0.74
	2	12.21	1.20	1.20	18.23	0.83	0.65
	5	11.59	0.75	0.75	15.50	0.98	0.47
Average		9.03	0.80	0.80	13.91	0.87	0.74

Table 4 **Average Relative Gaps When Transportation Times Are Exponential**

		(Det., Exp.)			(Exp., Exp.)		
		$\Delta FCFS$	ΔSP	ΔMR	$\Delta FCFS$	ΔSP	ΔMR
h_1	0.5	3.66	2.65	2.65	7.81	2.80	2.80
	1	5.29	0.64	0.64	10.05	0.65	0.62
$1/\theta_i$	2	7.47	1.49	1.49	12.76	1.12	1.08
	6	1.48	1.80	1.80	5.10	2.34	2.34
c_1/c_2	2	1.56	2.67	2.67	2.84	2.83	2.83
	5	4.42	1.47	1.47	8.94	1.58	1.58
	10	7.45	0.80	0.80	15.01	0.78	0.73
ρ	0.5	0.82	0.48	0.48	1.64	0.55	0.55
	0.8	3.36	1.67	1.67	7.59	1.98	1.97
	0.9	9.25	2.78	2.78	17.55	2.65	2.62
λ_1/λ_2	0.2	3.50	0.12	0.12	5.26	0.26	0.26
	0.5	3.95	0.70	0.70	7.86	0.99	0.99
	1	4.54	1.91	1.91	10.12	2.03	2.03
	2	5.23	2.85	2.85	11.49	2.89	2.89
	5	5.15	2.65	2.65	9.90	2.47	2.39
Average		4.48	1.64	1.64	8.93	1.73	1.71

Below we make several comments and suggestions based on the detailed numerical results; overall, we find that prioritization has value when it is done properly—as with the GMR.

1. In contrast to the centralized settings where the MR and SP polices outperform the FCFS policy (see, e.g., ABB), in some cases the FCFS policy outperforms the MR policies. For example, in Table 4 with $1/\theta_i = 6$, deterministic production and exponential transportation times, the average relative gap of the FCFS policy is less than the average relative gaps of the MR policies.

2. The MR policy, which is known to be optimal in the centralized M/M/1 make-to-stock system, is often identical to the SP policy in the two-echelon inventory system. In the 720 cases considered, there are only 17 cases in which the MR policy outperforms the SP policy. In these 17 cases, the production times are exponential and the relative gap between the FCFS and GMR policies is high. The reason is that since there is a delay between the warehouse and the DCs, it is more beneficial to keep the inventory at the high-priority DC instead of keeping products at the warehouse for this DC. In other words, if we want to keep inventory to decrease the chance of backlogging at the high-priority DC, it is more beneficial to keep it at the DC rather than at the warehouse.

3. As h_1 increases, the average relative gap of the FCFS policy increases, whereas the average relative gaps of the SP and MR polices decrease. This is expected

because as the base-stock level at DC 1 decreases, the probability of backlog at this DC increases. This observation suggests that prioritization is more important when inventory holding costs are high.

4. As the transportation times increase, $1/\theta_i$, the average relative gap of the FCFS policy decreases, and the average relative gaps of the SP and MR policies increase. Intuitively, as the transportation times increase, the two-echelon system behaves more like a decentralized system, and therefore the SP and MR policies introduced for the centralized systems become less effective.

5. As c_1/c_2 increases, the average relative gap of the FCFS policy increases, whereas the average relative gaps of the SP and MR polices decrease. Intuitively, prioritization has a higher value when the difference between backlog costs is higher. Interestingly, from the detailed results, we see that the average rate at which the gap of the FCFS policy increases when c_1/c_2 increases independently of the transportation times.

6. As ρ increases, the average relative gap of all three policies increases. Intuitively, as ρ increases, the average number of orders in the system increases, and therefore prioritization becomes more important. Interestingly, the rate at which the relative gaps of the SP and MR priority policies increase is low compared to the rate at which the gap of the FCFS policy increases.

7. The effect of λ_1/λ_2 on the average relative gaps is interesting. These gaps first increase and then decrease. This suggests that prioritization is more important when the demands made by different DCs are similar. Furthermore, our detailed numerical results demonstrate that when the utilization of the system is low, the rate at which the gap of the FCFS changes by changing λ_1/λ_2 is much lower than when the utilization of the system is high.

6.4. Comparing the GMR Policy with LQF and Myopic (T) Policies

In this section, we consider the LQF and myopic (T) policies and investigate the relative performance of the GMR policy compared to these policies. We recall that myopic (T) uses more information on the number of units at the DCs, thus, it is expected to lead to a lower cost. To examine the impact of the order crossing in time when the transportation times are stochastic, we also consider the GMROC policy. In §6.4.1 we investigate the performance of the GMR policy in a system with no order crossing by considering deterministic transportation time. To examine the performance of the GMR policy in a system with order crossing, we consider stochastic (exponential) transportation times in §6.4.2.

We present numerical examples for only 48 cases in Tables 5 and 6. In these cases, we set the mean production $\mu = 1$, the holding costs at the warehouse and DCs $h_0 = h_1 = h_2 = 0.5$, the backlog cost at DC 1 $c_1 = 10$, the manufacturer's utilization $\rho = \lambda/\mu = 0.8$, and the other parameters are varied as follows: the proportion of arrival rates $\lambda_1/\lambda_2 \in \{0.5, 1, 2\}$, the proportion of backlog costs $c_1/c_2 \in \{2, 5\}$, and the mean of transportation times $1/\theta_1 = 1/\theta_2 \in \{2, 6\}$. For each policy and test, we calculated the relative gap of this policy with respect to the GMR policy, as in (27). We use LQF and MT to denote the LQF and myopic (T) policies, respectively, in Tables 5 and 6.

Table 5 GMR Policy vs. Myopic (T), and LQF Policies When Transportation Times Are Deterministic

			(Det., Det.)		(Exp., Det.)	
$1/\theta_j$	c_1/c_2	λ_1/λ_2	ΔMT	ΔLQF	ΔMT	ΔLQF
2	2	0.5	0.46	−0.85	2.04	6.23
		1	3.80	−2.68	2.46	3.89
		2	5.29	−1.66	−1.29	−0.82
	5	0.5	0.02	10.50	−0.70	17.63
		1	0.24	4.17	1.16	16.69
		2	1.81	7.02	−2.30	6.59
6	2	0.5	3.03	−1.08	4.01	3.72
		1	5.09	−3.60	6.21	2.56
		2	4.67	−2.74	3.68	−2.37
	5	0.5	−0.11	4.00	0.47	11.24
		1	2.18	0.93	2.29	10.27
		2	1.51	−0.67	−0.50	7.64
Average			2.33	1.11	1.46	6.94

6.4.1. Performance of the GMR Policy in a System Without Order Crossing. In this section, we focus on systems with deterministic transportation times in which there is no order crossing. The results in Table 5 indicate that the GMR policy outperforms the myopic (T) and LQF policies in the majority of cases. This implies that when the assumption of the no order crossing holds, the GMR policy performs well.

6.4.2. Performance of the GMR Policy in a System with Order Crossing. In this section, we consider systems with exponential transportation times in which orders may cross each other in time, and compare the performance of the LQF, myopic (T), and GMROC with the GMR policy.

The results in Table 6 show that the myopic (T) and LQF policies outperform the GMR policy in most cases where transportation times are exponential, irrespective of the production times. Interestingly, $\Delta GMROC$, the relative gap between the GMR policy with and without order crossing, shows that the impact of the order crossing can be significant with an average of −20.23 in the examples that we considered. Moreover, although either myopic (T) or the LQF policy outperforms the GMROC policy, neither one of them dominates the GMROC policy.

Note that in practice, uncertainty in transportation times is typically much lower than the uncertainty in production and waiting times. The intiution behind

Table 6 GMR Policy vs. GMROC, Myopic (T), and LQF Policies When Transportation Times Are Exponential

			(Det., Exp.)			(Exp., Exp.)		
$1/\theta_j$	c_1/c_2	λ_1/λ_2	$\Delta GMROC$	ΔMT	ΔLQF	$\Delta GMROC$	ΔMT	ΔLQF
2	2	0.5	−10.52	−10.74	−13.42	−4.00	−4.00	−3.70
		1	−10.22	−3.45	−14.15	−5.03	−1.34	−5.92
		2	−11.05	−2.86	−15.36	−3.88	−5.19	1.36
	5	0.5	−11.83	−13.49	−7.41	−6.89	−8.26	8.69
		1	−10.26	−8.24	−12.34	−8.28	−8.65	3.95
		2	−12.25	−9.01	−13.49	−6.75	−9.03	−2.89
6	2	0.5	−32.16	−36.78	−40.77	−20.54	−13.08	−31.35
		1	−29.56	−35.74	−42.25	−21.78	−18.42	−26.29
		2	−27.06	−30.61	−41.34	−20.84	−11.53	−30.47
	5	0.5	−30.16	−38.33	−37.19	−23.24	−27.50	−19.91
		1	−29.38	−38.09	−41.04	−24.09	−19.64	−26.31
		2	−28.30	−41.95	−44.30	−22.13	−18.67	−31.01
Average			−20.23	−22.44	−26.92	−13.95	−12.11	−13.66

the benefit of the GMR in this setting is that this policy prioritizes the DCs at the production facility; therefore, its prioritization does not improve control of transportation times (see comment 4 in §6.3.2).

6.5. Comparing the GMR Policy with the Optimal Policy

In this section, we compare the optimal policy with the FCFS, MR, SP, GMR, GMROC, myopic (T), and LQF policies. We present numerical examples for only six cases in Table 7. In these cases, we set the mean production $\mu = 1$; the holding costs $h_0 = 0.5$, $h_1 = 1$, $h_2 = 0.5$; the backlog cost at DC 1 $c_1 = 10$; the proportion of arrival rates $\lambda_1/\lambda_2 = 1$; the mean of transportation times $1/\theta_1 = 1/\theta_2 = 2$ (transportation times are exponentially distributed); and the other parameters are varied as follows: the manufacturer's utilization $\rho = \lambda/\mu \in \{0.5, 0.6\}$, and the proportion of backlog costs $c_1/c_2 \in \{2, 5, 10\}$. For each policy we calculate the relative gap of this policy with respect to the optimal policy

$$\Delta^{\text{opt}}\bullet := \frac{C^{\bullet} - C^{\text{optimal}}}{C^{\text{optimal}}} \times 100. \tag{28}$$

The numerical results given in Table 7 demonstrate that if we ignore order crossing during transportation, the gap between the optimal policy and the GMR policy is high. But the relative gap between the optimal policy and the GMROC policy that considers order crossing is low. Moreover, from Table 7, and as demonstrated in §6.4.2, when we allow orders crossing, the performance of the GMROC policy is comparable with the LQF and mypoic (T) policies. Thus, although we developed the GMR policy for the no order crossing case, where it is analytically tractable, this policy appears quite effective even when order crossing is allowed. However, in such cases, similar to other policies, finding the optimal control of the GMROC policy is time consuming.

7. Conclusion

In this paper, we studied a two-echelon inventory system with several DCs where the supplier has limited production capacity and where both production and transportation times are general. We demonstrated that the queueing decomposition approach can be used to provide an exact analysis of several different policies for this system, namely, the FCFS, SP, MR, and GMR policies. We numerically compared the total cost of the system under these policies to the LQF, myopic (T), and optimal policies. This comparison shows that the GMR policy is beneficial in many cases. We also obtained the optimal cost of the system for several cases and demonstrated that the impact of order crossing can be high. Our derivations and numerical results suggest several insights on how to manage multiechelon systems with several classes of customers: (1) developing

Table 7 Comparison Between the Optimal Policy and the Other Policies

ρ	c_1/c_2	$\Delta^{\text{opt}} FCFS$	$\Delta^{\text{opt}} SP$	$\Delta^{\text{opt}} MR$	$\Delta^{\text{opt}} GMR$	$\Delta^{\text{opt}} GMROC$	$\Delta^{\text{opt}} MT$	$\Delta^{\text{opt}} LQF$
0.5	2	22.56	21.34	21.34	20.80	0.02	3.03	3.45
	5	22.82	18.28	18.28	18.28	0.54	1.26	5.03
	10	24.82	17.91	17.91	17.91	4.05	0.15	7.62
0.6	2	24.26	22.22	22.22	22.19	2.95	3.07	5.31
	5	26.96	21.82	21.82	21.56	0.21	1.32	11.02
	10	29.40	21.58	21.58	21.40	0.44	0.69	13.70
Average		25.14	20.52	20.52	20.35	1.37	1.59	7.69

new prioritization policies for the multiechelon system such as the GMR policy; (2) prioritizing according to the backlog costs when the uncertainty is in production times rather than transportation times is effective; and (3) in contrast to the centralized setting, in the decentralized setting prioritization is more valuable only when the warehouse is out of stock.

Finally, we note that the model presented in the paper can be applied to spare systems with two transit legs, one to move the replacement part from the warehouse to the DC, and other to move the failed unit from the DC to the warehouse for repair and refurbishment. With regard to the time required to move the failed unit from the DC to the warehouse, such additional transportation time can be incorporated into our model as follows: Assume, as is typical in manufacturing models of make-to-stock queues, that supply of raw materials is ample. Then, we model this transportation time for facility j as an independent $M/G_j/\infty$, the output process of such a queue—that is the arrival process to the warehouse from DC j—is still Poisson and so is the total arrival process to the warehouse. Therefore, we can consider the transportation time in the paper as the convolution of the transportation time from the DC to the warehouse with the one from the warehouse to the DC and all results in the paper would still hold. Although assuming ample raw materials in spare part systems may not be realistic, we still believe that the priority policies introduced in this paper could be useful in practice.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2014.0494.

Acknowledgments

The authors thank the editors and referees for their valuable comments. This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grants to the authors.

References

Abate J, Whitt W (1992) Numerical inversion of probability generating functions. *Oper. Res. Lett.* 12(4):245–251.

Abouee-Mehrizi H, Baron O (2014) Queue decomposition and its applications in state-dependent queues. Working paper, University of Waterloo, Ontario.

Abouee-Mehrizi H, Balcıoğlu B, Baron O (2012) Strategies for a centralized single product multiclass $M/G/1$ make-to-stock queue. *Oper. Res.* 60(4):803–812.

Abouee-Mehrizi H, Baron O, Berman O (2013) The exact analysis of the GMR policy in a multiechelon inventory system with m DCs. Working paper, University of Waterloo, Ontario.

Abouee-Mehrizi H, Berman O, Shavandi H, Zare AG (2011) An exact analysis of a joint production-inventory problem in two-echelon inventory systems. *Naval Res. Logist.* 58(8):713–730.

Axsater S (1990) Simple solution procedures for a class of two-echelon inventory problems. *Oper. Res.* 38(1):64–69.

Axsater S (2006) *Inventory Control*, 2nd ed. (Springer, New York).

Baron O (2008) Regulated random walks and the LCFS backlog probability: Analysis and applications. *Oper. Res.* 56(2):471–486.

Benjaafar S, ElHafsi M, Huang T (2010) Optimal control of a production-inventory system with both backorders and lost sales. *Naval Res. Logist.* 57(3):252–265.

Bertsimas D, Nakazato D (1995) The distributional Little's law and its applications. *Oper. Res.* 43(2):298–310.

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

de Véricourt F, Karaesmen F, Dallery Y (2000) Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies. *Oper. Res.* 48(5):811–819.

de Véricourt F, Karaesmen F, Dallery Y (2001) Assessing the benefits of different stock-allocation policies for a make-to-stock production system. *Manufacturing Service Oper. Management* 3(2):105–121.

de Véricourt F, Karaesmen F, Dallery Y (2002) Optimal stock allocation for a capacitated supply system. *Management Sci.* 48(11):1486–1501.

Gayon J, de Véricourt F, Karaesmen F, Dallery Y (2009) Stock rationing in an $M/E_r/1$ multiclass make-to-stock queue with backorders. *IIE Trans.* 41(12):1096–1109.

Glasserman P (1997) Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* 45(2):244–257.

Graves SC (1985) A multiechelon inventory model for a repairable item with one-for-one replenishment. *Management Sci.* 31(10):1247–1256.

Ha A (1997a) Stock-rationing policy for a make-to-stock production system with two priority classes and backordering. *Naval Res. Logist.* 44(5):457–472.

Ha A (1997b) Optimal dynamic scheduling policy for a make-to-stock production system. *Oper. Res.* 45(1):42–53.

Jing X, Lewis M (2011) Stockouts in online retailing. *J. Marketing Res.* 48(2):342–354.

Levi R, Roundy R, Shmoys D, Sviridenko M (2008) A constant approximation algorithm for the one-warehouse multiretailer problem. *Management Sci.* 54(4):763–776.

Mak HY, Shen ZJ (2009) A two-echelon inventory-location problem with service considerations. *Naval Res. Logist.* 56(8):730–744.

Muharremoglu A, Tsitsiklis JN (2008) A single-unit decomposition approach to multiechelon inventory systems. *Oper. Res.* 56(5):1089–1103.

Parker RP, Kapuscinski R (2004) Optimal policies for a capacitated two-echelon inventory system. *Oper. Res.* 52(5):739–755.

Peña Perez A, Zipkin P (1997) Dynamic scheduling rules for a multiproduct make-to-stock queue. *Oper. Res.* 45(6):919–930.

Roundy RO, Muckstadt JA (2000) Heuristic computation of periodic-review base stock inventory policies. *Management Sci.* 46(1):104–109.

Schneeweiss C, Schroder H (1992) Planning and scheduling the repair shops of the Deutsche Lufthansa AG: A hierarchical approach. *Production Oper. Management* 1(1):22–33.

Sennott L (1999) *Stochastic Dynamic Programming and the Control of Queueing Systems* (John Wiley & Sons, Hoboken, NJ).

Sherbrooke CC (1968) METRIC: A multi-echelon technique for recoverable item control. *Oper. Res.* 16(1):122–141.

Svoronos A, Zipkin P (1991) Evaluation of one-for-one replenishment policies for multiechelon inventory systems. *Management Sci.* 37(1):68–83.

Veatch MH, Wein LM (1994) Optimal control of a two-station tandem production/inventory system. *Oper. Res.* 42(2):337–350.

Wein LM (1992) Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.* 40(4):724–735.

Zheng Y, Zipkin P (1990) A queueing model to analyze the value of centralized inventory information. *Oper. Res.* 38(2):296–307.