POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 23, No. 1, January 2014, pp. 110–126
ISSN 1059-1478 | EISSN 1937-5956 | 14 | 2301 | 0110
DOI 10.1111/poms.12041

Coordinated Logistics: Joint Replenishment with Capacitated Transportation for a Supply Chain

Nasuh C. Büyükkaramikli
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, PO Box 513, 5600MB, Eindhoven, The Netherlands, n.c.buyukkaramikli@tue.nl

Ülkü Gürler
Department of Industrial Engineering, Bilkent University, Ankara, Turkey, ulku@bilkent.edu.tr

Osman Alp
Department of Industrial Engineering, TED University, Ankara, Turkey, osman.alp@tedu.edu.tr

In this study, we consider the integrated inventory replenishment and transportation operations in a supply chain where the orders placed by the downstream retailer are dispatched by the upstream warehouse via an in-house fleet of limited size. We first consider the single-item single-echelon case where the retailer operates with a quantity based replenishment policy, (r,Q), and the warehouse is an ample supplier. We model the transportation operations as a queueing system and derive the operating characteristics of the system in exact terms. We extend this basic model to a two-echelon supply chain where the warehouse employs a base-stock policy. The departure process of the warehouse is characterized in distribution, which is then approximated by an Erlang arrival process by matching the first two moments for the analysis of the transportation queueing system. The operating characteristics and the expected cost rate are derived. An extension of this system to multiple retailers is also discussed. Numerical results are presented to illustrate the performance and the sensitivity of the models and the value of coordinating inventory and transportation operations.

Key words: joint replenishment; transportation; inventory; logistics
History: Received: July 2010; Accepted: October 2012 by Jayashankar Swaminathan, after 2 revisions.

1. Introduction

In this study, we jointly consider the inventory replenishment and transportation operations in a supply chain with stochastic demand. Our work has been motivated by the current practices as well as the existing gap in the literature regarding the coordination of the stock control and dispatch operations in supply chains. As illustrated by our numerical findings, simultaneous consideration of inventory and transportation management functions raises interesting issues and provides managerial insights such as the significance of the joint consideration of the replenishment and the transportation functions, optimal fleet sizes, and the impact of delays due to transportation unit.

To reflect the significance of the issue, we note that the total logistics activities comprise approximately 1.28 trillion USD or about 8.5% of the US GDP in 2011 (Burnson 2012). Two major components of the logistics costs are the transportation costs and inventory carrying costs where transportation (largely trucking costs) accounted for 63% while inventory carrying costs accounted for 33% in the US economy in 2002 (FHWA 2005). The sheer size of the expenses involved is an incentive for both shippers and carriers to find ways to reduce them. Better management of the physical assets for transporting goods and also of inventories themselves may provide significant savings (http://www.smartops.com). The integrated management of transportation capacity and inventory becomes especially crucial in developing economies where the total logistics supply chain costs account for about 24% of the GDP, of which about 10% is due to indirect costs such as inefficient logistics activities resulting in higher inventories and shortages (Dobberstein et al. 2005). In a typical developed market, such indirect costs account for about 5% of the GDP. Moreover, truck driver shortages have become a major logistical concern in developed countries (RT 2010), causing truck unavailability.

In a supply chain, one of the most commonly used mode of dispatching the orders is in-house transportation. In-house transportation has the advantage of providing more controlled and reliable transportation

together with increased visibility of the products in transit. Furthermore, in certain environments, specifically designed vehicles are needed; for example, hazardous materials or cold chains for fresh food or medical supplies require custom-designed vehicles with specific temperature and humidity controls. Although in-house transportation is commonly used and is an essential stage in the fulfillment of the customer orders, joint modeling of inventory replenishment and transportation dynamics and investigation of the impacts and the restrictions faced by each function have not been much elaborated in the literature.

In supply chains where these logistics activities are not coordinated, inventory and transportation operations are managed separately with different and possibly conflicting objectives. In particular, the inventory manager searches for the "optimal" inventory control parameters that would minimize inventory related costs (holding, backordering, and ordering) whereas the transportation manager searches for the least number of trucks sufficient to yield acceptable congestion levels, utilization ratios, and minimum fleet related costs. Evidently, uncoordinated decisions might not yield optimal operating characteristics for the whole system, as these two logistics activities are closely interrelated. Regarding the fleet size issue, if the decision makers do not adopt a coordinated perspective, they might fail to correctly assess the overall cost of operating the system. If an over-estimated fleet size is used, the delays due to transportation are reduced but the operating costs of the transportation unit will be inflated. On the other hand, if a smaller fleet than the optimal is used, although the business would keep going, impact of delays due to transportation capacity would have a negative impact on inventory management practices. Our work provides insights regarding the optimal choice of the fleet size and also the additional costs that would be incurred if it is set in a sub-optimal way.

In this study, we address the joint modeling of replenishment and transportation functions in a supply chain. To introduce the settings and the main issues, we start with a single-echelon model with a single retailer, stochastic demand, and capacitated in-house fleet. This problem results in a classical inventory problem with random lead times, where the lead times have the special distribution induced by the underlying queueing system at the transportation unit. We derive the exact expressions for the operating characteristics and the long term expected cost function when the retailer faces unit Poisson demand. In the second model, we extend the model to a two-echelon supply chain with single supplier and single retailer where the retailer employs an (r,Q) policy and the warehouse employs a base stock inventory policy. In this setting, the warehouse faces orders of size Q that may not be satisfied immediately due to insufficient stocks. In this setting the departure process of the orders from the warehouse, which constitutes the arrival process of the transportation unit, is different than the arrival process to the warehouse. The inter-departure times of this process is characterized in probability distribution, which is then used to propose an approximation by an Erlang process with matching first two moments. This approximation enables the analysis of the underlying queuing systems for transportation operations. The corresponding (approximate) operating characteristics and the expected cost rate are established for the two-echelon model. As a further extension, the applicability of the model to N retailers is demonstrated where the retailers adopt a joint (Q,S) replenishment policy.

Before summarizing our findings, let us first briefly refer to the related literature to better locate the contributions of our study. The existing studies on coordinated replenishment and distribution problems in supply chains mostly consider these problems separately, and the settings arising from their integration have not been explored in detail. To the best of our knowledge, there are only three studies that consider the impact of cargo capacity on coordinated inventory replenishment decisions in a stochastic demand environment. The first one, which also motivated our work to a large extent, is by Cachon (2001), and the other two are more recent extensions by Gurbuz et al. (2007) and Tanrikulu et al. (2010). Cachon (2001) and Tanrikulu et al. (2010) analyze a supply chain environment where the joint replenishment orders of retailers are dispatched by an ample supplier with capacitated trucks. It is assumed that fleet size is unlimited. Gurbuz et al. (2007) also assume an ample supplier; however, the joint orders are shipped by a single truck from the warehouse to a cross-dock facility. If the joint order size exceeds the truck capacity, excess quantity is still shipped with an additional penalty cost. In all of these studies, the capacity constraints are either explicitly or implicitly on the size of an individual order; they have no limitation on the number of orders in transit any time due to fleet size restriction. This assumption has two implications—one practical and one theoretical. In practice, such limitations do exist. A supply chain that has opted to have its own fleet (of finite size) may hesitate to utilize third parties due to delivery quality concerns or the administrative burden of emergency management. From a theoretical perspective, the assumption of unlimited fleet size implies that the orders are dispatched immediately so long as the warehouse has

stock. That is, the stochastic delivery delays encountered by the lower echelon are only a function of the inventory control dynamics at the upper echelon. Limited fleet size introduces another source of delays that has not been studied before.

Another stream of research also addresses truck cargo capacity under different settings for single location inventory systems (e.g., Alp et al. 2003, Ernst and Pyke 1993, Toptal et al. 2003, Yano and Gerchak 1989), and also in the context of inventory/routing problems (e.g., Ball et al. 1983, Federgruen and Zipkin 1984, Sindhuchao et al. 2005, Tanrikulu et al. 2010, Toptal and Cetinkaya 2006). The methodology in these papers differs greatly from ours due to either the fleet size limitation considered herein and/or the stochastic nature of demand in our model. We therefore do not further elaborate such literature.

Finally, we note that the joint inventory replenishment problem has been studied extensively in literature, in both deterministic and stochastic environments, and the early works go back to Balintfy (1964), who developed the continuous-review *can-order* policy, and Ignall (1969), who is the first to study the optimal joint replenishment policy. The optimal policy, even for two items and zero lead times, has a very complicated structure. Hence most of the existing studies focus on intuitive heuristic policy classes. Related works include Renberg and Planche (1967) who first proposed the (Q,**S**) policy, Pantumsinchai (1992), who presented an exact analysis of this policy under Poisson demands, and Cheung and Lee (2002), Nielsen and Larsen (2005), and Ozkaya et al. (2005, 2006).

Our study makes a number of contributions in theoretical and application aspects. From a theoretical perspective, our main contribution lies in providing a unified modeling framework to integrate the stochastic dynamics of inventory replenishment and transportation operations. Our approach rests on characterizing in distribution the departure process of the warehouse. This departure process becomes the arrival process of the transportation unit, which carries the items to replenish the retailer's inventory. For a single-echelon environment, we extend the classical (r,Q) model in a way that integrates the transportation and inventory replenishment operations and provide exact expressions for the total expected cost function. For a two-echelon environment, we provide an approximate total expected cost function that utilizes the inter-departure distribution. We list the special cases where our approach becomes an exact analysis for the two-echelon system. We also obtained an interesting result that states that the variance of inter-departure times from the warehouse is bounded above by the variance of the inter-arrival times of the joint orders from the retailers.

From a practical perspective, our analysis and numerical results provide several managerial insights. First of all, we show that there is a considerable value in coordinating the transportation and inventory operations. In section 5, we show that system-wide cost of an uncoordinated system might be 175% higher than the cost of the coordinated system for a particular problem instance. We illustrate that explicitly modeling the limited size of the available fleet has a significant impact on the resulting system costs, and that the cost inefficiency can be as high as 68% if the fleet size limitations are ignored. We identify the minimum and maximum fleet size thresholds where congestion levels are permissible and operations are economical, respectively. We believe that such benchmarks would be of use for investment and/or supply chain design decisions. In our numerical study, we observe that diminishing marginal return of increased transportation capacity does not necessarily hold in general, and that under certain settings, insights gained with unlimited fleet size do not match with those when fleet size is limited. Furthermore, we address the characteristics of the environments where using the upper echelon as cross-dock may be more beneficial.

The rest of the study is organized as follows: in section 2, we describe the problem environment in detail and analyze the single-echelon version of the problem. In section 3, we extend our analysis to a two-echelon environment. In section 4, we further extend our models to a multiple retailers, single warehouse environment. In section 5, numerical experiments and observations are provided. Finally in section 6, an overall summary of the study and future research directions are provided.

2. Coordinated Logistics: Single Echelon

Consider a continuous review, single-item, single-echelon inventory system with an ample warehouse and a retailer. The retailer faces stationary and independent unit Poisson demand with rate λ, and unmet demands are fully backordered. Holding and shortage costs incurred at the retailer are denoted by h and b, respectively, per unit per time. The retailer operates with an (r,Q) policy where an order of size Q is placed whenever the inventory position at the retailer drops to r. The warehouse operates with an in-house fleet of K trucks that are utilized for delivering the orders placed by the retailer. The orders received by the warehouse are immediately processed and streamed to the transportation unit for dispatching. There is a cost of $\phi(K,C)$ to maintain a fleet of K trucks where each of them can

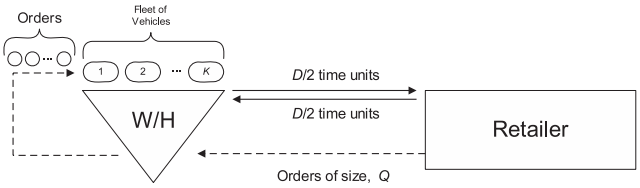
accommodate C units. This cost component includes costs for maintenance, repair, depreciation, cost of truck drivers and the apprentice for loading and unloading items, etc.

For each truck utilized for order shipments, a fixed cost of $A(C)$ is incurred independent of the quantity loaded in the truck. One of the components of this fixed cost would be the fuel cost of transportation, which is a function of the truck capacity. We assume that at least 50% truck utilization is attained for order deliveries. Enforcing a minimum truck utilization is a common practice in industry due to transportation limitations and large fixed costs, as well as environmental regulations that encourage the reduction of carbon dioxide emissions by several means. Moreover, we also restrict the order size Q to be less than a full truck load, C, as this has several benefits. First of all, $Q > C$ implies delaying the shipment and would not be desirable when the unit backordering cost is higher than the unit holding cost (see Cachon 2001). Moreover, if order integrality is adapted, $Q > C$ would result in a shipment delay of a full truck for which a fixed cost is charged anyway. Therefore, when the unit backordering cost is higher than the unit holding cost, we do not expect the optimal Q being larger than C. Apart from this intuitive justification for the assumption that $Q \leq C$, we should note that if Q is allowed to be greater than C, then the operating system should adopt more complicated protocols regarding order integrity and order sequencing. In particular, if $Q > C$, several trucks may be needed to carry an order, and at an instant only a portion of the order may be available at the inventory. In such cases whether or not order integrality should be adopted or whether the full trucks should be dispatched right away or wait until the whole order is ready become important issues both practically and theoretically. No matter which procedure is selected for implementation, the resulting system would clearly be more complicated. Note also that the optimal dispatching procedures under such situations are not known. Consequently, we assume that $C/2 < Q \leq C$.

The duration of a one-way trip from the warehouse to the retailer is given by $D/2$ where D is the duration of a return trip. Consequently, the replenishment lead time for the retailer is $L = D/2$ if there is an available truck at the transportation unit at the order instant and is larger than $D/2$ if there is a delay due to truck unavailability. Figure 1 depicts a representation of the system under consideration.

As the warehouse is an ample supplier, the orders received by the warehouse are immediately processed and relayed to the transportation unit for shipment on a first-come-first-served basis, without any delay. The time between each successive retailer order has

Figure 1 Representation of the System



an Erlang distribution with shape and scale parameters Q and λ, respectively, as the retailer observes unit Poisson demand with rate λ and employs an (r,Q) policy. Under these settings, the transportation unit operates as an $E_Q/D/K$ queue where the arrival process to the queue is the departure process of retailer orders from the warehouse, the deterministic service time D corresponds to the fixed transit times of vehicles, and the number of servers is the fleet size, K.

2.1. Waiting Time Distribution of Retailer Orders

In this section, we characterize the random waiting time, W_q, of an order at the transportation unit that operates as an $E_Q/D/K$ queue. Note that the effective replenishment lead time for the retailer is a random variable and is given by $L = D/2 + W_q$.

The waiting time distribution in a multi-server $GI/D/c$ queue is equivalent to that of a single-server $GI^*/D/1$ queue, where the inter-arrival time distribution GI^* is the convolution of c inter-arrival times with distribution GI (Tijms 1995, p. 321), that is, shorter inter-arrival times are compensated with a higher number of servers, yielding stochastically equivalent waiting times. In our setting, this implies that the waiting time distribution of an $E_Q/D/K$ queue is identical to that of a $M/D/Q\times K$ queue. To find the waiting time distribution of an $M/D/c$ queue where $c = Q\times K$, we adopt the method of Franx (2001). In order to be coherent with the common terminology, we use *customer* and *server* for a joint order and a truck, respectively.

Let p_i denote the stationary probability that there are i customers in the system, given as

$$p_i = \sum_{j=0}^{c} p_j \frac{(\lambda D)^i}{i!} e^{-\lambda D} + \sum_{j=c+1}^{i+c} p_j \frac{(\lambda D)^{i+c-j}}{(i+c-j)!} e^{-\lambda D}, \quad i \in \mathcal{N}.$$

The p_i's constitute the solution of an infinite system of linear equations subject to the normalization $\sum_{i=0}^{\infty} p_i = 1$. According to Tijms (1995, p. 289), the state probabilities p_j of an $M/D/c$ queue exhibit the geometric tail property, $p_j \approx \delta\gamma^{-j}$ for large j, where $\gamma \in (1,\infty)$ is the unique solution of $\lambda D(1-\gamma) + c \ln(\gamma) = 0$ and δ is given by $\delta = (c - \lambda D\gamma)^{-1} \sum_{i=0}^{c-1} p_i(\gamma^i - \gamma^c)$. Through this geometric tail property, the infinite system of linear equations for the p_j's is

reduced to a finite system by replacing p_j by $p_M(1/\gamma)^{j-M}$ for $j > M$ for an appropriately chosen M. Let q_i be the stationary probability that the queue contains i customers, where $q_0 = \sum_{i=0}^{c} p_i$ and $q_i = p_{i+c}$ for $i > 0$. Also, define the cumulative probability that there are j or less customers in the queue as $G_j = \sum_{i=0}^{j} q_i$. Then, referring to Franx (2001), the distribution of the waiting time (W_q in our case) in the queue of a $M/D/c$ system is given as

$$F_{W_q}(w) = e^{-\lambda(a_w D - w)} \sum_{j=0}^{a_w c-1} G_{a_w c-j-1} \frac{\lambda^j (a_w D - w)^j}{j!}, \quad (1)$$

where a_w is the greatest integer less than or equal to $\frac{w}{D} + 1$ for $w \geq 0$. This is a mixed distribution with discrete and continuous parts. Observe that $F_{W_q}(w)$ implicitly depends on K as the number of servers is $c = K \times Q$.

2.2. Inventory Related Costs at the Retailer

We next derive the expected holding and backordering cost rates incurred at the retailer conditioned on a given value of W_q, by following the approach of Axsäter (1990). This approach is based on the observation that a unit ordered by the retailer is used to fill the $(r + Q)$th subsequent demand following this order. Recall that the lead time is given by $L = D/2 + W_q$ and the retailer employs an (r,Q) policy. Let $S = r + Q$, and $l = D/2 + w$ is a given effective lead time for a particular realization of $W_q = w$. Then, the expected holding and backordering costs per unit per time, $g(S|l_i)$, is given as

$$g(S|l) = \frac{1}{\lambda}[S(h+b)F_P(S,\lambda l) - \lambda l(h+b)F_P(S-1,\lambda l) + b(\lambda l - S)], \quad (2)$$

where $F_P(y,\lambda l)$ denotes the cumulative probability distribution of a Poisson variable with rate λl (Cachon 2001). When $Q = 1$, a unit demand always triggers an order. However for $Q > 1$, the demands arriving at the retailer wait until a total of Q units accumulate, and only after this is an order placed. Suppose a demand arrives at the retailer at time τ, but a replenishment decision is delayed until time $\tau + t$. That order is supplied to the retailer at $\tau + t + l$. Let M denote the total number of demand arrivals at the retailer between τ and $\tau + t$. When $M = m$, the unit demand that occurred at τ is used to fill the $(r + Q - m)$th subsequent demand after $\tau + t$. It is known that M has a discrete uniform distribution on $0,\ldots,Q - 1$ (see Axsäter 1993). Hence, the expected holding and backordering cost per time per unit for the retailer with a given effective lead-time l is

$$\frac{1}{Q}\sum_{m=0}^{Q-1} g(r + Q - m|l), \quad (3)$$

where the function g is given by Equation (2).

2.3. Policy Optimization

For the ample supplier, taking the expectation of the cost expression in Equation (3) with respect to the distribution of $W = W_q$, we can write the expected cost rate of the system as

$$AC(r,Q,K) = \lambda\frac{A(C)}{Q} + \phi(K,C) + \lambda\int_w \frac{1}{Q}\sum_{m=0}^{Q-1} g(r+Q-m|D/2+w)dF_{W_q}(w). \quad (4)$$

Hence, the following optimization problem is to be solved to find the optimal policy parameters, r,Q, and K:

$$\min_{r;Q:C/2 < Q \leq C;K} AC(r,Q,K).$$

The expected unit holding and backorder cost rate given by Eqution (3) is convex in r (see Axsäter 1993). As expectation is a linear operator, $AC(r,Q,K)$ is also convex in r. Therefore, the optimal re-order point $r^*(Q,K)$ can be found by a convex optimization algorithm for given Q and K values. However, as the total cost rate $AC(r^*(Q,K),Q,K)$ is not necessarily convex in Q, the optimal shipment quantity Q^* for given K is obtained by a complete search over the feasible interval $(C/2,C]$. Although unimodality over K is observed in our numerical analysis, an exhaustive search for the optimal K value is needed as there is no analytical result in this respect. However, there are natural lower and upper bounds on the value of K for a given Q. The total cost rate for the system is finite only if the underlying queue satisfies the stability condition $\rho = \frac{\lambda \times D}{K \times Q} < 1$. This means that there is a minimum number of trucks that is needed for the queueing system to reach the steady state for a given Q. Let $K_{\min}(Q)$ be the smallest positive K that satisfies $\rho < 1$. In our numerical experiments, we observe that for a fixed value of Q, each truck added to $K_{\min}(Q)$ brings a diminishing decrease in total expected holding and backordering costs. Hence, there is a sufficiently large K value that approximates the total inventory costs of an unlimited fleet size situation. As it is natural to expect $\phi(K,C)$ to be an increasing function in K, that value of K would be an upper bound on the optimal value of the fleet size.

3. Coordinated Logistics: Two-Echelon Supply Chain

In this section, we extend our analysis to a two-echelon inventory system. In this setting the warehouse is no longer an ample supplier but employs a base-stock policy, i.e., whenever an order of size Q is received from the retailer, an order of the same size is placed immediately. This policy can be represented by $(S_w - Q, S_w)$ where the order-up-to level S_w represents the inventory position of the warehouse at any given time. Note as a convention that, if the inventory level of the warehouse is less than $S_w - Q$ as an initial condition, then an immediate order that is greater than Q is placed so that the inventory position is raised up to S_w at the start of the planning horizon. Retailer orders are satisfied on a first-come-first-served basis and the integrity of the orders is sustained. As partial shipments are not allowed, the optimal order-up-to level will be an integer multiple of the batch size, i.e., $S_w = \Delta \times Q$, where Δ is a nonnegative integer. L_w denotes the replenishment lead time between the warehouse and its ample supplier whereas A_w denotes the fixed cost of ordering at the warehouse.

Contrary to the single-echelon model, a retailer order may not be relayed to the transportation unit immediately due to possible stock-out occasions at the warehouse. This creates another source of delay for the retailer order. Because of this, the departure process of the orders at the retailer is no longer Erlang. Therefore, the queueing system that governs the transportation unit becomes a general $G/D/K$ queue. Consequently, the effective replenishment lead time for the retailer becomes $L = D/2 + W_q + W_s$ where W_s denotes the random variable for the delay due to lack of sufficient inventory at the warehouse and W_q is the waiting time at the transportation unit in a $G/D/K$ queue.

3.1. Departure Process of Retailer Orders at the Warehouse

In this part, we derive the probabilistic characteristics of the departure process of the orders at the warehouse that employs a $(S_w - Q, S_w)$ policy. Note that the $(S_w - Q, S_w)$ policy is equivalent to a base-stock $(S - 1,S)$ policy when a batch of size Q is considered as a single unit and S is set to Δ in terms of the new unit that corresponds to one batch. For simplicity, we derive the expressions for an $(S - 1,S)$ system in this section.

Consider a warehouse operating under an $(S - 1,S)$ policy where $S \geq 0$. Suppose the warehouse faces unit demands with i.i.d. inter-arrival times given by $\{X_j, j \geq 1\}$, and probability density function (pdf) and cumulative distribution functions (cdf) $f_X(\cdot)$ and $F_X(\cdot)$, respectively. We set $\sum_{j=m}^{n} a_j = 0$ if $m > n$ for any $a_j \in \Re$ without loss of generality. Let $X_0 = 0$, and $f_{X^{(j)}}(\cdot)$ and $F_{X^{(j)}}(\cdot)$ denote the pdf and cdf of jth arrival time, $X^{(j)} = \sum_{n=1}^{j} X_n$, respectively. First, suppose that a demand arrives at time τ that immediately triggers an order. Due to the nature of the $(S - 1,S)$ policy, this order satisfies the Sth subsequent demand whose arrival time is $\tau + \sum_{n=1}^{S} X_n$. Whenever this demand arrives, if the warehouse has positive on-hand inventory, then this demand is immediately dispatched. Hence, its departure time from the warehouse would be $\tau + \sum_{n=1}^{S} X_n$, the same as its arrival time. Otherwise, it waits for the arrival of the triggered order, and its departure time will be $\tau + L_w$ (see Axsäter 1990 for more details). Letting DT_S denote the departure time of the Sth subsequent demand after τ, we have

$$DT_S = \tau + \max\left(\sum_{n=1}^{S} X_n, L_w\right).$$

Next, consider the jth demand that arrives after τ, which triggers another order and arrives at time $\tau + \sum_{n=1}^{j} X_n + L_w$. Then, we can write the departure time DT_{j+S} of the $(j + S)$th subsequent demand after τ as

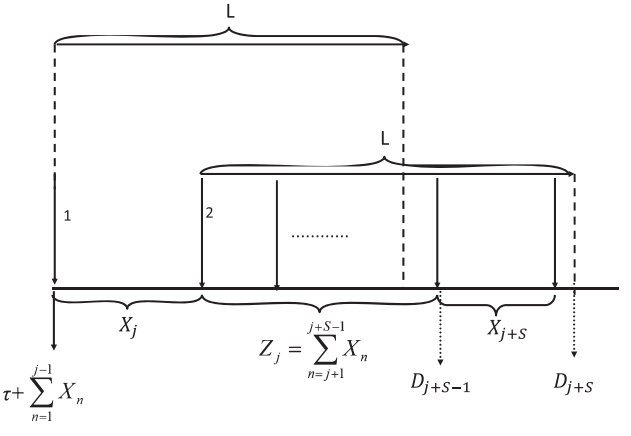
$$DT_{j+S} = \tau + \max\left(\sum_{n=1}^{j+S} X_n, \sum_{n=1}^{j} X_n + L_w\right).$$

Let Y_{j+S} be the time between the departures of the jth and $(j - 1)$st demands after τ:

$$\begin{aligned} Y_{j+S} &= DT_{j+S} - DT_{j+S-1} \\ &= \max\left(\sum_{n=j}^{j+S} X_n, X_j + L_w\right) - \max\left(\sum_{n=j}^{j+S-1} X_n, L_w\right). \end{aligned}$$

An illustration of the consecutive departures from warehouse is given in Figure 2. Let $Z_j = \sum_{n=j+1}^{j+S-1} X_n$.

Figure 2 Illustration of the Consecutive Demand Departures



Then the cdf $F_Z(\cdot)$ of Z_j is identical to $F_{X^{(S-1)}}(\cdot)$, the cdf of $X^{(S-1)}$ for all j.

The following expression for Y_{j+S} provides a more convenient representation:

$$Y_{j+S} = \begin{cases} X_{j+S} & \text{if} \quad (L_w - Z_j) \le \min(X_j, X_{j+S}) \\ X_j + Z_j + X_{j+S} - L_w & \text{if} \quad X_j < (L_w - Z_j) \le X_{j+S} \\ L_w - Z_j & \text{if} \quad X_{j+S} < (L_w - Z_j) \le X_j \\ X_j & \text{if} \quad (L_w - Z_j) > \max(X_j, X_{j+S}) \end{cases}. \tag{5}$$

From the above construction, we observe that inter-departure times of the orders from the warehouse have identical distributions and $Y_{j+S} = X_{j+S}$ if $S = 0$. The distribution function of these identical variables, say Y, is denoted by $F_Y(y)$ and is given below.

THEOREM 1. *The distribution function $F_Y(y) = F_{Y_j}(y)$ of the inter-departure time Y_j of an (S−1,S) inventory system with deterministic lead time and unit renewal demands is identical for all j, which is given as follows:*

$$F_Y(y) = \begin{cases} F_X(y) \int_{L_w - y}^{\infty} \bar{F}_X(L_w - z)\, dF_Z(z) \\ \quad + \int_0^{L_w} \int_0^{\min(y, L_w - z)} F_X(L_w - z + y - x_2)\, dF_X(x_2)\, dF_Z(z) & \text{if } S > 0 \\ F_X(y) & \text{if } S = 0 \end{cases} \tag{6}$$

for $y \ge 0$, $L_w > 0$, and $\bar{F}_X(\cdot) = 1 - F_X(\cdot)$.

PROOF. See the Appendix.

As Y_j's are identical variables, $E[Y_j] = E[Y]$ and $\mathrm{Var}[Y_j] = \mathrm{Var}[Y]$ for all j. On the other hand, again from the definition of the inter-departure time given by Equation (5), we observe that the departure process has $(S + 1)$-dependence, as each departure depends on the $(S + 1)$ preceding arrivals. We provide the mean and the variance of the inter-departure times below.

THEOREM 2. *Let E[X] and Var[X] be the expectation and the variance of the inter-arrival time X, respectively. Then, E[Y] = E[X], Var[Y] ≤ Var[X], where*

$$Var[X] - Var[Y] = 2 \int_{z=0}^{\infty} \left\{ \left(\int_{x=L_w - z}^{\infty} \bar{F}_X(x)\, dx \right) \left(\int_{x=0}^{L_w - z} F_X(x)\, dx \right) \right\} dF_Z(z).$$

PROOF. See the Appendix.

The above result indicates that (i) the mean inter-arrival times to the warehouse are the same as the mean inter-departure times from the warehouse, which is expected in order to have a stable system, and (ii) the variance of the inter-departure times from the warehouse is no more than the variance of the inter-arrival times. This implies that the lead time at the warehouse has a "smoothing" effect to reduce the variability of the inter-departure times, which is not obvious immediately. An explanation can be as follows: due to the $(S - 1,S)$ policy employed at the warehouse, if an inter-arrival is too short $(X < L)$, then it is likely that the later demand will wait until the end of the lead time, extending the corresponding inter-departure time of the orders $(Y \ge X + L)$. Similarly, if an inter-arrival is too long $(X > L)$, then the corresponding inter-departure could be shorter if the former order waits for stock availability $(Y \le X - L)$. Hence, extreme inter-arrival times may be pulled down or pushed up to moderate inter-departure intervals, resulting in possible variance reduction.

3.2. Approximations for System Analysis

In this section, we present two approximations that we have used for the analysis of the two-echelon system. The first one is related to the departure process of the warehouse and the second is about the independence of W_s and W_q. The departure process of the warehouse characterized by Theorem

1 introduces a challenging problem in terms of the queuing system at the transportation unit, as the arrival process has identical but serially correlated inter-arrival times, which renders it impossible to identify the waiting time distribution at the transportation unit explicitly. To the best of our knowledge, even the waiting time distribution of $G/D/K$ queues with independent renewal arrivals is a difficult problem (see, e.g., Schleyer and Furmans 2007, Whitt 1993). We make the observation from Equation (5) that when the lead time at the warehouse is zero or if it tends to infinity, the departure process of the warehouse coincides with that of the arrivals, resulting in Erlang departures. For moderate values of the lead time, the departure process is not an exact Erlang process, but it does not deviate significantly from that either, as illustrated in section 5. Therefore, in order to overcome the difficulty, we approximate the exact inter-departure distribution by a suitable Erlang distribution. Note that approximating some random characteristics with an Erlang distribution is also a commonly used approach in the literature (see, e.g., Altiok 1985, Bitran and Tirupati 1988, Graves 1985, Whitt 1982). Therefore, we propose to use an Erlang departure process whose first two moments match with those of the true departure process (which is characterized by Theorems 1 and 2) for an approximate system analysis. As a convention, if the shape parameter from matching the moments results in a non-integer value, we use the closest integer.

The total waiting time of a retailer order before being shipped with a truck is given by $W = W_s + W_q$. The common stochastic dynamics underlying the realizations of W_s and W_q may impose a dependency between the variables W_s and W_q, which seems to be non-trivial to identify exactly. In order to characterize the probability distribution of W, one can either refer to simulation methods or to some approximate analytical methods. For practical purposes, we propose an approximation that is based on the assumption that W_s and W_q are independent random variables. In particular, we assume that an order departing from the warehouse and arriving at the transportation unit finds the transportation system in the steady state and the waiting times at the warehouse and at the transportation unit are independent. It will be established in the numerical section that the performance of the analytical model under an Erlang approximation and an independence assumption deviates from the simulated system by a negligible amount unless the traffic is highly congested.

Consequently, the waiting time distribution of a retailer order at the warehouse is approximated by

$$F_W(x) = \int_{y=0}^{x} F_{W_q}(x-y)\, dF_{W_s}(y), \tag{7}$$

where $F_{W_q}(x)$ is given by Equation (1). For characterization of the distribution function $F_{W_s}(\tau)$ of the random delay W_s at the warehouse, we refer to Ozkaya et al. (2005) who provide the delay distributions at the upper echelon for various stochastic joint replenishment policies. When the retailer employs the (r,Q) policy and the warehouse order-up-to level is S_w, it is given by

$$F_{W_s}(\tau) = \begin{cases} 0 & \tau < 0 \\ 1 - F_E(L_w - \tau, \Delta \times Q, \lambda) & 0 \le \tau \le L_w, \\ 1 & \tau \ge L_w \end{cases} \tag{8}$$

where $F_E(x,k,\lambda)$ denotes the distribution function of an Erlang random variable with shape and scale parameters k and λ and with density $f(x,k,\lambda)$.

3.3. Policy Optimization

In this section, we derive expressions for the total relevant expected costs per unit time achieved at the steady state, under the approximations explained above. We verified through simulations that the system converges to a steady state under any arbitrary starting inventory level as far as the traffic ratio is less than one. As expected, the convergence rate depends on the problem parameters, in particular on the traffic ratio and the starting inventory level.

For a particular realization of the retailer lead time, $L = l$, the expected holding and backordering costs incurred at the retailer side are still given by Equation (3) as explained in section 2.2. Consequently, the expected holding and backordering costs incurred at the retailer are obtained by un-conditioning this expression as follows:

$$U(r,Q,K,\Delta) = \int_{w} \frac{1}{Q} \sum_{m=0}^{Q-1} g(r+Q-m|D/2+w)\, dF_W(w),$$

where F_W is given by Equation (7). Note that $W = W_q + W_s$, and $F_W(w)$ implicitly depends on K and $S_w = \Delta \times Q$, which in turn affect the $F_{W_q}(\cdot)$ and $F_{W_s}(\cdot)$.

Next, we consider the holding cost rate incurred at the warehouse. We use a method similar to that of Axsäter (1990) and note that a holding cost for a joint retailer demand of size Q that arrives at the warehouse at time τ is incurred if the Δth subsequent joint retailer demand arrives after $\tau + L_w$. Hence the expected time a retailer order incurs a holding cost at the warehouse inventory is

$$\int_{x=L_w}^{\infty} (x - L_w) f_E(x, \triangle Q, \lambda) dx,$$

where $f_E(x,\Delta Q, \lambda)$ denotes the Erlang pdf with parameters ΔQ and λ. This expression reduces to

$$\frac{\triangle Q}{\lambda} F_P(\triangle Q, \lambda L_w) - L_w F_P(\triangle Q - 1, \lambda L_w).$$

In addition to the holding time at the warehouse, holding cost is also incurred while waiting for an available truck to be dispatched at the transportation unit. Let $E[W_q]$ be the expected dispatching waiting time at the transportation unit. Then the holding cost incurred at the warehouse level, $WH(Q,\Delta,K)$ is given by

$$WH(Q, \Delta, K) = h_w Q \frac{\lambda}{Q} \left\{ E[W_q] + \frac{\Delta Q}{\lambda} F_P(\Delta Q, \lambda L_w) - L_w F_P(\Delta Q - 1, \lambda L_w) \right\}. \quad (9)$$

Then the expected cost rate of the entire supply chain is given by

$$AC(r, Q, K, \Delta) = \lambda \frac{A(C) + A_w}{Q} + \phi(K, C) + \lambda U(r, Q, K, \Delta) + WH(Q, \Delta, K). \quad (10)$$

The first part of this expression represents the retailer and warehouse order setup cost rates. The other parts represent the fleet maintenance costs, holding and backorder costs incurred at the retailer level, and the holding cost incurred at the warehouse level, respectively. Considering the truck utilization constraint, the optimization problem is stated as

$$\min_{r, Q \in (C/2, C], K, \Delta} AC(r, Q, K, \Delta).$$

4. Extensions

4.1. *N*-Retailers

The analysis discussed in the previous sections can be extended to a system with N retailers that use a joint replenishment policy where the joint order size is fixed as Q. As the (Q,S) joint replenishment policy exhibits this structure and is a simple and commonly used one, we illustrate how to extend our models under this policy. Suppose the N retailers are supplied by a single warehouse. Let λ_i denote the demand rate of retailer i and $\lambda_0 = \sum_{i=1}^{N} \lambda_i$. Also let $L_i = D/2 + l_i$ be the total time required to replenish retailer i after a loaded truck departs from the warehouse.

4.1.1. Single Echelon. Similar to the single retailer case, when $Q = 1$, a unit demand always triggers a joint replenishment order. However for $Q > 1$, the demands arriving at retailers wait until a total of Q units accumulate and then a joint order is placed. Suppose a demand arrives at retailer i at time τ, but a joint replenishment decision is delayed until time $\tau + t$. That joint order is supplied to the retailer at $\tau + t + l$. Let M_i denote the total number of demand arrivals for retailer i between τ and $\tau + t$. When $M_i = m_i$, the unit demand that occurred at τ is used to fill the $(S_i - m_i)$th subsequent demand after $\tau + t$. Let $M_0 \geq M_i$ be the total number of retailer demands (including i) that have occurred in $(\tau,\tau + t]$. When $M_0 = m_0$, the probability that m_i of these demands are from retailer i is binomial with parameters m_0 and success probability $r_i = \lambda_i/\lambda_0$. In accordance with the single retailer case, the total expected inventory holding and backordering cost can be given as

$$U_s(Q, \mathbf{S}, K) = \int_w \frac{1}{Q} \sum_{m_0=0}^{Q-1} \sum_{m_i=0}^{m_0} \binom{m_0}{m_i} (r_i)^{m_i} (1 - r_i)^{m_0 - m_i} g_i(S_i - m_i \mid L_i + w)\, dF_{W_q}(w).$$

Consequently, the total expected cost function can be written as follows:

$$AC(Q, \mathbf{S}, K) = \lambda_0 \frac{A(C)}{Q} + \phi(K, C) + \sum_{i=1}^{N} \lambda_i U_s(Q, \mathbf{S}, K). \quad (11)$$

4.1.2. Two-Echelon. For the two-echelon system, similar to the above discussion, the expected inventory holding and backordering costs for each retailer i can be written as follows:

$$U(Q, \mathbf{S}, K, \Delta)_i = \int_w \frac{1}{Q} \sum_{m_0=0}^{Q-1} \sum_{m_i=0}^{m_0} \binom{m_0}{m_i} (r_i)^{m_i} \times (1 - r_i)^{m_0 - m_i} g_i(S_i - m_i | L_i + w)\, dF_W(w).$$

Hence, the total expected operating cost for the entire supply chain is

$$AC(Q,\mathbf{S},K,\Delta)=\lambda_0\frac{A(C)+A_w}{Q}+\phi(K,C)+\sum_{i=1}^{N}\lambda_i U(Q,\mathbf{S},K,\Delta)_i+WH(Q,\Delta,K), \quad (12)$$

where $WH(Q,\Delta,K)$ is as in Equation (9) with the exception that λ is replaced by λ_0.

It can easily be verified that the expressions in Equations (11) and (12) reduce to expressions in Equations (4) and (10) for $N = 1$, respectively. Given the total expected cost functions in Equations (11) and (12), optimal policy parameters can be sought in accordance with the single retailer case as explained in sections 2.3 and 3.3.

4.2. Time-Based Policies

The (Q,S) policy employed by the retailers is an example of a "quantity-based policy" in which an inventory replenishment is triggered by the accumulation of demand quantity. As an alternative, a "time-based policy" can be considered, in which the inventory replenishments are triggered by accumulation of a certain time. A commonly implemented periodic review time-based policy is the (**R**, T) policy where T is the length of the period and **R** is the vector of order-up-to levels. The inventory position of the ith retailer is raised up to R_i at every T time units. The order quantity placed at the end of a period is the total demand observed during that period, which is a Poisson random variable with expected value $\lambda_0 T$. Hence under this policy the order quantity is not a constant but a randomly changing quantity, which also implies that the arrivals to the transportation unit are not constant either. Therefore, when the system operates with an in-house fleet for transportation, a modeling approach similar to the one employed in our study cannot be adopted directly under the (**R**, T) policy. To be more specific, note that an order of random size arrives to the truck queue in every T time units. As the number of trucks available at any instance is at most K, it is possible that an arriving order must be split and partially shipped at different times, with more than one truck. Similarly it is possible that two or more orders are consolidated and shipped on the same truck. Due to such complications, optimal dispatching protocols should also be sought, and the focus of the problem changes. We therefore keep time-based policies out of our scope.

However, detailed comparative analyses of quantity-based and time-based policies are available in the literature. See, for example, Cetinkaya et al. (2006) and Mutlu et al. (2010), who both report that quantity-based policies are superior to the time-based policies in an environment where the shipments are consolidated at the warehouse, which serves multiple retailers that employ a joint replenishment policy. Another related study is by Shang et al. (2010), who compare the (**R**,T) policy to a quantity-based policy under a serial multi-echelon inventory system. The authors show that the overall system benefits from switching from a time-based policy to a quantity-based policy.

5. Numerical Study

In this section, we report the results of our numerical tests conducted to gain insights on different aspects of the problem under concern. As a test bed, all combinations of the following parameters are used unless otherwise stated: $N \in \{1,2,16\}$, $\lambda_0 \in \{4,8,16,32\}$, $h = 1$, $b \in \{4,8,16,32\}$, $C \in \{4,8,12,16,32\}$, $D \in \{2,4,8\}$. It is assumed that the retailers are identical in their demand rate, holding and backordering costs, and lead times for the multiple retailers case. We also assume that the fixed retail order cost has the following structure: $A(C) = \alpha \times C^{\gamma}$ for $\alpha,\gamma > 0$ and $\phi(K,C) = KC^{\beta}$. Our numerical study focuses on investigating mainly three issues: (i) the value of coordination and the importance of modeling the limited fleet, (ii) the impact of a limited fleet on the integrated system of inventory and transportation, and (iii) the accuracy of the Erlang approximations. These issues are discussed in the following sections. Although all these issues deserve a detailed discussion, due to space restrictions we sometimes restrict our attention to special cases and sometimes report our findings without providing the actual numerical results. Further details can be obtained from the authors.

5.1. Analysis of a Limited Fleet

In this part, we analyze the value of coordinating transportation and inventory decisions and the impact of problem parameters on the operating characteristics. We take $A(C) = \alpha C^{\gamma}$ with $\alpha \in \{0.25,1,4\}$ and $\gamma = 1$, similar to the experimental set of Cachon (2001), who analyzes the unlimited fleet version of this problem. We do not associate any cost to fleet maintenance in this section to mainly focus on inventory ordering related costs whenever necessary.

5.1.1. Value of Coordination. Consider a problem instance where the warehouse is an ample supplier, $\lambda = 8, N = 1, b = 8, C = 16, D = 8, A(C) = 0.25C, \phi(C,K) = KC^{0.5}$. For the coordinated system, the optimal parameters for this problem instance are $Q^* = 16, S^* = 49$, and $K^* = 5$, whereas the optimal cost is 34.64. Assume that this system is operated in an uncoordinated fashion. If the inventory manager

assumes that the transportation unit will deliver the order in exactly $D = 8$ time units without any delay (note that this assumption corresponds to assuming infinite fleet size), then she will prefer to operate with $Q = 11$ and $S = 45$. In this case, the minimum number of trucks to operate the system with $\rho < 1$ is six. If the transportation unit decides to operate with six trucks, then the realized system-wide costs would be 95.28, which is 175.03% higher than the cost of the coordinated system (see Table 1). We define this percentage as the "value of coordination." In such a case, the traffic ratio will be 0.97. Due to this high congestion, orders will wait at the warehouse for truck availability for excessive times and hence the delivery lead time will be much higher than eight on the average in practice. To alleviate this high congestion, if both units negotiate to operate with more trucks, system-wide costs get lower, but there is still a considerable value for coordinating the system. In particular, if the transportation unit operates with seven, eight, or nine trucks, then the value of coordination becomes 22.64%, 33.29%, and 44.82%, respectively (see Table 1). In the uncoordinated system, the inventory manager might anticipate the congestion due to truck unavailability and she might inflate the value of D while determining the optimal policy parameters. If D is inflated by 50% and set to 12, then the inventory manager will obtain $Q = 12$ and $S = 63$. In this case, $K_{\min} = 9$. In such a case, there is still considerable room for improvement. In particular, the value of coordination becomes 19.43%, 25.62%, 34.49%, and 43.45% if the transportation unit operates with 9, 10, 11, or 12 trucks, respectively. Table 1 also summarizes the results if the delivery lead time is inflated by different percentages.

5.1.2. Importance of Modeling Limited Fleet. We next comment on the importance of explicitly modeling the limited size of the available fleet. To assess this, we evaluated the increase in the operating costs that would have been incurred if the optimal policy parameters of an unlimited fleet size model are used, when the system is in fact operated with a limited number of trucks. We conduct our experiments for a single-echelon system in this part in order to base our comparisons on exact cost figures. In particular, let (Q^*, S^*) be the optimal parameters of the model with limited fleet size (see section 2) and (Q_U^*, S_U^*) be those of unlimited fleet size model. Also let $K_{min} = K_{min}(Q^*)$ be the minimum number of trucks needed to satisfy the stability condition, $\rho < 1$. To assess the contribution of our model, we compare the cost that would be incurred if the system is operated with (possibly) sub-optimal parameters (Q_U^*, S_U^*) to the optimal cost rates for a fleet size ranging from K_{min} to $K_{min} + 3$. We present the results in Table 2, where the percentage losses in the expected cost rate are given and ∞ indicates that the Q_U^* and the given fleet size result in a $\rho > 1$ that violates the stability condition. From the table, we observe that the cost inefficiency could be as high as 68%, and the loss decreases when the number of available trucks or the capacity of the trucks increase, as expected intuitively. These findings confirm that our modeling

Table 1 Value of Coordination for a Problem Instance

				$AC(Q,S,K)$				Value of Coordination (%)			
D	Q	S	$K_{\min}$	$K_{\min}$	$K_{\min}+1$	$K_{\min}+2$	$K_{\min}+3$	$K_{\min}$	$K_{\min}+1$	$K_{\min}+2$	$K_{\min}+3$
8	11	45	6	95.28	42.49	46.18	50.17	175.03	22.64	33.29	44.82
10	12	54	7	64.28	47.43	51.19	55.19	61.95	19.49	28.97	39.04
12	12	63	9	53.37	56.13	60.1	64.1	19.43	25.62	34.49	43.45

Table 2 Percentage Losses under (Q_U^*, S_U^*) of the Unlimited Fleet Case, $N = 4$, $D = 8$

		$b = 16$				$b = 32$			
λ_0	C	$K_{\min}$	$K_{\min}+1$	$K_{\min}+2$	$K_{\min}+3$	$K_{\min}$	$K_{\min}+1$	$K_{\min}+2$	$K_{\min}+3$
16	2	54.95	25.24	8.68	2.58	68.04	38.84	17.17	6.96
	4	27.65	3.74	0	0	41.23	9.22	2.31	0.52
	8	2.13	0	0	0	6.22	0	0	0
	16	0	0	0	0	∞	4.95	0.10	0
	32	∞	47.65	0	0	∞	∞	4.44	0
32	2	44.93	26.97	14.95	8.16	56.46	36.06	20.46	10.91
	4	41.40	10.61	1.73	0	57.53	22.52	7.26	2.42
	8	13.82	1.09	0	0	19.26	1.35	0	0
	16	1.55	0	0	0	1.92	0	0	0
	32	0.84	0	0	0	7.42	0	0	0

approach provides an opportunity for significant cost savings.

5.1.3. Sensitivity of Optimal Policy to the Problem Parameters. The computational results presented here are based on the approximations of section 3.2. Our main findings can be summarized as follows:

(i) The optimal order size, Q^*, optimal truck utilization, Q^*/C, and optimal order-up-to levels, S^*, are non-increasing in the fleet size, K (see Table 3).

(ii) There is an upper bound on the fleet size that yields operating characteristics practically equivalent to an unlimited fleet environment. We observe that this upper bound is non-increasing in fixed ordering cost. As lower fixed costs lead to lower Q^* values with more frequent shipments, the system becomes more vulnerable to operating under limited fleet sizes and requires more trucks to behave like an unlimited fleet environment.

(iii) For a given fleet size K, Q^* and S^* are non-decreasing in the truck capacity C.

(iv) Q^* and hence the truck utilization are non-increasing in the backordering cost.

(v) An increase in the transit time, D, or the demand rate, λ, increases the traffic ratio, ρ, as well as the expected demand during lead time. This leads to larger Q^* values and larger truck utilization.

(vi) From Table 3, it is observed that the supplier operates as a cross-dock facility for small fleet sizes with $\Delta^* = 0$. This can be explained with the observation that for small K, higher Q values are needed, which results in large inventory carrying costs and hence the system pulls down the value of Δ.

(vii) As K increases, we do not always observe a diminishing return in the optimal cost rate when operated with optimal (Q^*, S^*) values (see Figure 3 for an example). This example indicates that the marginal benefit of additional truck capacity does not necessarily decrease as this capacity increases, which is contrary to what is commonly observed in the literature for capacitated problems (see, e.g., Alp and Tan 2008).

(viii) No monotonic relation is observed between N and Q^*. The total demand rate λ_0 is fixed and the total order-up-to levels are observed to decrease as N gets smaller. This is because the total demand is partitioned to fewer retailers, reducing the uncertainty. For an unlimited fleet size, Cachon (2001) states that Q^* is always non-increasing in N in single-echelon environments, as larger order sizes in the (Q,S) policy lead to larger variations among the inventory levels of retailers as N increases and this brings elevated operating costs. However, under limited fleet size, especially with a scarcity of trucks, a reduction in Q increases the traffic ratio, which in turn has negative impacts on the holding and backordering costs due to increased delays. This negative impact may dominate the impact of the increase in the number of retailers, and hence there are cases where Q^* (and the truck utilization) increases as N increases with limited fleet size (see Table 4).

Figure 3 Illustration of the Change of $AC(Q^*, (S^*(Q^*, K)), K, C)$ in K

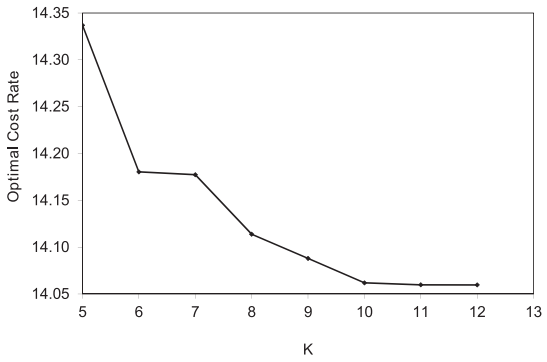


Table 3 The Effects of the Change in $A(C)$ and K on Total Cost Rate when $N = 1$, $\lambda_0 = 4$, $D = 8$, $L_w = 2$, and $b_i = 32$ for all i

	$A(C)$								
	$\alpha = 0.25, \gamma = 1$			$\alpha = 1, \gamma = 1$			$\alpha = 4, \gamma = 1$		
K	$(Q^*, S^*, \triangle^*, C^*)$	$E[W_q]$	$E[W_s]$	$(Q^*, S^*, \triangle^*, C^*)$	$E[W_q]$	$E[W_s]$	$(Q^*, S^*, \triangle^*, C^*)$	$E[W_q]$	$E[W_s]$
2	(21,49,0,32)	0.03	2	(22,49,0,32)	0.01	2	(32,58,0,32)	0	2
3	(14,43,0,16)	0.03	2	(16,44,0,16)	0.00	2	(16,44,0,16)	0	2
4	(11,40,0,16)	0.01	2	(15,43,0,16)	0.00	2	(16,44,0,16)	0	2
9	(8,30,1,8)	0	0.28	(8,30,1,8)	0	0.28	(8,30,1,8)	0	0.28
14	(5,28,2,4)	0	0.11	(5,28,2,4)	0	0.11	(5,28,2,4)	0	0.11

For each problem instance, the best truck capacity, C^*, is found by searching from the set {2,4,8,16,32}.

(ix) Finally we elaborate on the waiting times due to insufficient inventory and unavailability of a truck for dispatching. Table 3 also reports the expected waiting time in the transportation queue, $E[W_q]$, and in the warehouse $E[W_s]$. In general, an increase in Q leads to a decrease in $E[W_q]$. However, $E[W_s]$ is decreasing in Q only when Δ is fixed. It is observed that when the supplier operates as a cross-dock, $E[W_s]$ is significantly larger than $E[W_q]$.

Table 5 provides instances to observe the impact of W_q and W_s in a two-echelon environment. For $L_w = 2$, an increase in Δ from 0 to 1 reduces the total waiting times drastically (from 5.27 to 3.29) by reducing the operating costs at the retailers more than the increased inventory costs at the warehouse. On the other hand, when warehouse lead time $L_w = 1$, W_q dominates W_s, and keeping inventory at the warehouse does not bring further benefits.

Table 4 Impact of Number of Retailers, N, when $\lambda_0 = 4$, $b_i = 4$, $\alpha = 1$, $\gamma = 1$, and $D = 8$ in Single-Echelon Environments

K	C	$N = 1$ (Q^*, S_i^*)	$N = 2$ (Q^*, S_i^*)	$N = 4$ (Q^*, S_i^*)	$N = 16$ (Q^*, S_i^*)
2	32	(30, 41)	(21, 17)	(21,9)	(23,3)
3	16	(15, 28)	(16, 15)	(16,8)	(14,2)
4	16	(15, 28)	(16, 15)	(15,8)	(11,2)

5.2. Accuracy of the Approximation

We first examine the accuracy of the Erlang approximation adopted in section 3, where the departure process of the warehouse operating under an $(S_w - Q, S_w)$ policy is approximated by an Erlang (Q', λ_0') process, whose first two moments match with the exact arrival process. Figure 4 illustrates two examples regarding the performance of this approximation. In this figure, exact and approximated distribution functions are depicted. When $L_w = 4$, the exact and approximated scale and shape parameters turned out to be the same, that is, $\lambda' = \lambda_0$ and $Q' = Q$, and the two cdfs are almost identical (Figure 4a). When $L_w = 6$, the adjusted parameters are $\lambda' = 5$ and shape $Q' = 5$, which are different from λ_0 and Q, but again the approximation is highly accurate (Figure 4b). For the lead time values beyond these limits, that is, for $L_w > 6$ and $L_w < 4$, the approximated scale and shape parameters were not changed, and the approximation performed perfectly. We see that Erlang approximation performs highly satisfactorily in terms of the distribution function.

Next, we examine the accuracy of the assumption that W_s and W_q are independent (see section 3.2). In order to assess the error due to this assumption, we obtained the exact and the approximate cdf of the waiting time $(W = W_s + W_q)$, the former obtained by a simulation study and the latter by Equations (6)–(8). This approximation is affected by

Table 5 Impact of L_w, $E[W_q]$, and $E[W_s]$ on Δ when $Q = 11$, $K = 3$, $\lambda_0 = 4$, $b = 32$, $h = 1$, $\alpha = 0.25$, $\gamma = 1$, and $C = 16$

	$\Delta = 0$			$\Delta = 1$		
L_w	$AC(\cdot)$	$E[W_q]$	$E[W_s]$	$AC(\cdot)$	$E[W_q]$	$E[W_s]$
2	73.40	3.27	2	71.54	3.23	0.06
1	72.88	3.27	1	73.45	3.27	0.00

Figure 4 Comparison of the Erlang Approximation and the Exact Queue Inter-arrival Times when $\lambda_0 = 4, Q = 4, \Delta = 5$; $L_w = 4$ in (a) and $L_w = 6$ in (b)

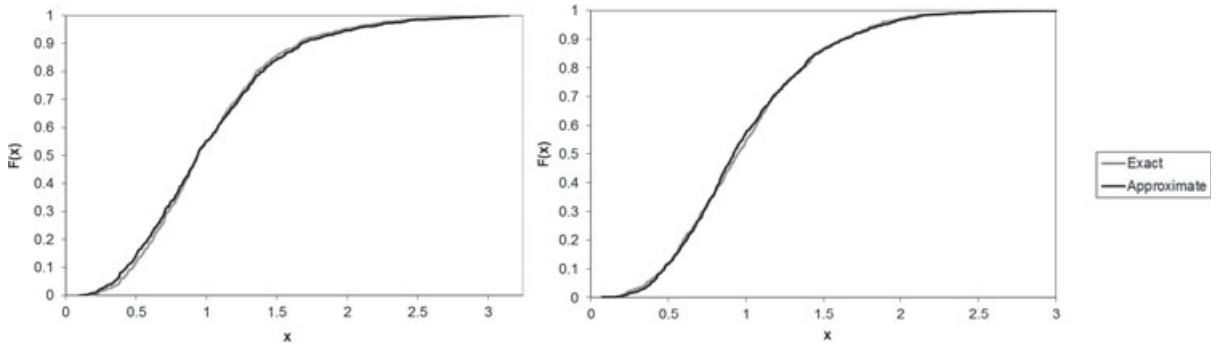


Figure 5 Comparison of Exact and Approximate W when $Q = 4$, $\lambda_0 = 8$, $D = 8$, $L_w = 3$, $K = 17$; $\Delta = 3$ in (a) and $\Delta = 6$ in (b)

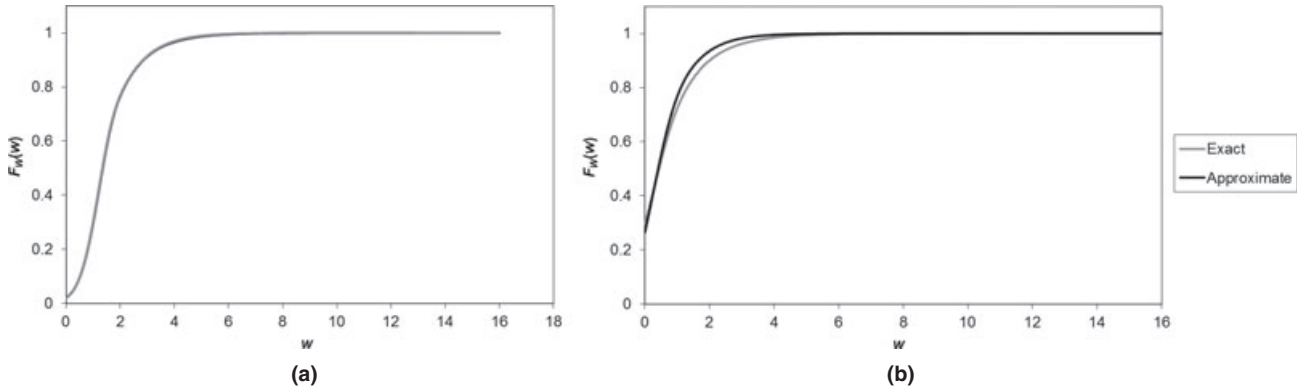


Table 6 The Accuracy of the Approximation for Different ρ

ρ	$\%\mu_{\lvert err\rvert}$	min % err	median % err	max % err
$\rho \leq 0.7$	0.026	0.000	0.021	0.088
$0.7 < \rho \leq 0.8$	0.205	0.002	0.132	0.766
$0.8 < \rho \leq 0.9$	0.534	0.025	0.432	1.885
$0.9 < \rho \leq 0.92$	0.618	0.002	0.288	3.702

the traffic ratio, which has a direct impact on W_q, the base stock level at the warehouse, $S_w = \Delta Q$, and the ratio $\frac{\Delta Q}{\lambda_0 L_w}$. This ratio indicates the number of orders that the base stock level can satisfy during the warehouse lead time. Figure 5 depicts interlaced graphs of the exact and approximate cdfs of W. As can be seen from the graph, when $\Delta = 3$ the functions overlap perfectly, but when $\Delta = 6$, there is some discrepancy. The maximum and the average differences between the functions are 0.0454 and 0.0083, respectively, when $\Delta = 6$. The same values are 0.0028 and 0.0005, respectively, when $\Delta = 3$.

We also assessed the impact of approximations on the operating characteristics of the system by comparing the cost rate of proposed analytical model to the cost rate obtained by simulating the real system. In simulations, we used a run length of 100,000 warehouse orderings and the average of 20 replications are taken as the expected cost rate of the system. The accuracy of the approximation is measured by $\%err = 100 \times (AC_{sim} - AC_{app})/AC_{sim}$, where AC_{sim} is the exact cost rate and AC_{app} is the approximate one. As an example, for the problem instance of Figure 5b, which corresponds to relatively large differences between the exact and approximate waiting time cdfs, the optimal fleet size turns out to be 19 and $\%err$ is 0.81, quite a small error. We also investigated 144 other scenarios, and Table 6 summarizes the accuracy of the approximations. The following parameter ranges are used to generate these problem instances: $L_w \in \{1,2\}, b \in \{4,16,32\}, \lambda \in \{8,16,32\}, C \in \{2,8,16,32\}, D = 8, N = 4$. We arbitrarily set Q,S, and Δ values so that 36 problem instances are generated for each of the traffic ratio ranges shown in Table 6 . We simulated the system until 1 million warehouse orders are generated and discarded the initial 30% of the simulation time.

In Table 6, we report the cases for $\rho \leq 0.92$. This is due to the fact that as the traffic intensity increases, assessment of the quality of approximation by simulations becomes less reliable, as the actual system not only converges to a steady state extremely slowly but also shows significant variance. Hence capturing the true performance by simulation becomes very difficult and making comparisons with the approximations would be misleading.

6. Conclusion and Future Studies

In this study, we considered the effect of transportation fleet capacity on the performance of a supply chain. We considered single- and two-echelon environments with a single retailer and a single warehouse. It is assumed that the retailer adopts the (r,Q) policy and the warehouse operates with a fleet of vehicles to satisfy the orders placed by the retailer. We derive the exact operating characteristics for the single-echelon environment and propose an approximate analysis for a two-echelon environment. Our results are also extended to a N-retailer environment.

Our results indicate that consideration of the transportation capacity and the fleet size in conjunction with the inventory replenishment operations can lead to substantial savings for the whole supply chain. We believe that our study may have important applications for supply chain design. In addition our study can be extended to contractual design agreements, especially for the 3PL provider firms that supply logistic service to retailer chains.

Appendix: Proofs

PROOF OF THEOREM 1. When $S = 0$, it is straightforward to observe that $Y_j = X_j$ from Equation (5) and hence $F_Y(y) = F_X(y)$. For $S > 0$, we rewrite Equation (5) in terms of the events $E_1 - E_4$ as

$$Y_j = \begin{cases} X_j & \text{if} \quad E_1 \equiv (L - Z_{j-S}) \le \min(X_{j-S}, X_j) \\ X_{j-S} + Z_{j-S} + X_j - L & \text{if} \quad E_2 \equiv X_{j-S} < (L - Z_{j-S}) \le X_j \\ L - Z_{j-S} & \text{if} \quad E_3 \equiv X_j < (L - Z_{j-S}) \le X_{j-S} \\ X_{j-S} & \text{if} \quad E_4 \equiv (L - Z_{j-S}) > \max(X_{j-S}, X_j). \end{cases}$$

Then, $F_{Y_j}(y) = P(Y_j \le y) = \sum_{n=1}^{4} P(Y_j \le y, E_n)$. As Y_j's have identical distribution, we let $Y_j \equiv Y$, with cdf F_Y, and as $X_j's$ are identical, we use F_X to denote their cdf and F_Z to denote that of Z. Note that

$$\begin{aligned}
P(Y \le y, E_1) &= P(X_j \le y, L - Z_{j-S} \le \min(X_{j-S}, X_j)) \\
&= \int_{z=L-y}^{\infty} \int_{x_2=L-z}^{y} \int_{x_1=L-z}^{\infty} dF_{X_1}(x_1)\, dF_{X_2}(x_2)\, dF_Z(z) \\
&= \int_{z=L-y}^{\infty} \bar{F}_X(L - z)\, F_X(y)\, dF_Z(z) - \int_{z=L-y}^{\infty} \bar{F}_X(L - z)\, F_X(L - z)\, dF_Z(z) \\
&\equiv 1.1 + 1.2 \\
P(Y \le y, E_2) &= P(X_{j-S} + Z_{j-S} + X_j - L \le y, X_{j-S} < L - Z_{j-S} \le X_j) \\
&= \int_{z=0}^{L} \int_{x_2=0}^{\min(y, L-z)} \int_{x_1=L-z}^{L-z+y-x_2} dF_{X_1}(x_1)\, dF_{X_2}(x_2)\, dF_Z(z) \\
&= \int_{z=0}^{L} \int_{x_2=0}^{\min(y, L-z)} F_X(L - z + y - x_2)\, dF_X(x_2)\, dF_Z(z) \\
&\quad - \int_{z=0}^{L} F_X(L - z)\, F_X(\min(y, L - z))\, dF_Z(z) \\
&\equiv 2.1 + 2.2 \\
P(Y \le y, E_3) &= P(L - Z_{j-S} \le y, X_j < L - Z_{j-S} \le X_{j-S}) \\
&= \int_{z=L-y}^{\infty} \bar{F}_X(L - z)\, F_X(L - z)\, dF_Z(z) \\
&\equiv 3.1 \\
P(Y \le y, E_4) &= P(X_{j-S} \le y, L - Z_{j-S} > \max(X_{j-S}, X_j)) \\
&= \int_{z=0}^{\infty} \int_{x_2=0}^{\min(y, L-z)} \int_{x_1=0}^{L-z} dF_{X_1}(x_1)\, dF_{X_2}(x_2)\, dF_Z(z) \\
&= \int_{z=0}^{L} F_X(L - z)\, F_X(\min(y, L - z))\, dF_Z(z) \\
&\equiv 4.1.
\end{aligned}$$

We observe that 1.2 cancels with 3.1, and 2.2 cancels with 4.1, yielding the result. □

PROOF OF THEOREM 2. As for a stable system we need $E[Y] = E[X]$, we skip the proof of this part, which is similar to the proof below. To show $\text{Var}[Y] \leq \text{Var}[X]$, we find $E[Y^2]$. Let $a = L - Z_{j-S}$. Then from Equation (1) we have

$$\begin{aligned}
E[Y^2] = &\int_{a=-\infty}^{L} \int_{x_2=a}^{\infty} \int_{x_1=a}^{\infty} x_2^2 f_Z(L-a)\, dF_{X_1}(x_1)\, dF_{X_2}(x_2)\, da \\
&+ \int_{a=-\infty}^{L} \int_{x_2=a}^{\infty} \int_{x_1=0}^{a} (x_2 + x_1 - a)^2 f_Z(L-a)\, dF_{X_1}(x_1)\, dF_{X_2}(x_2)\, da \\
&+ \int_{a=-\infty}^{L} \int_{x_2=0}^{a} \int_{x_1=a}^{\infty} a^2 f_Z(L-a)\, dF_{X_1}(x_1)\, dF_{X_2}(x_2)\, da \\
&+ \int_{a=-\infty}^{L} \int_{x_2=0}^{a} \int_{x_1=0}^{a} x_1^2 f_Z(L-a)\, dF_{X_1}(x_1)\, dF_{X_2}(x_2)\, da.
\end{aligned}$$

Evaluating the above integrals, we get

$$\begin{aligned}
E[Y^2] = E[X^2] \int_{a=-\infty}^{L} f_Z(L-a)\, da + \int_{a=-\infty}^{L} \Bigg\{ &2aF_X(a)\left(a\bar{F}_X(a) - \int_{x_2=a}^{\infty} x_2 dF_X(x_2)\right) \\
&+ 2\int_{x_1=0}^{a} x_1 dF_X(x_1)\left(\int_{x_2=a}^{\infty} x_2 dF_X(x_2) - a\bar{F}_X(a)\right)\Bigg\} f_Z(L-a)\, da
\end{aligned}$$

Letting $z = L - a$ we rewrite the above as

$$\begin{aligned}
E[Y^2] = E[X^2] + 2\int_{z=0}^{\infty} \Bigg\{ &\left(\int_{x=L-z}^{\infty} x dF_X(x) - (L-z)\bar{F}_X(L-z)\right) \\
&\cdot \left(\int_{x=0}^{L-z} x dF_X(x) - (L-z)F_X(L-z)\right)\Bigg\} dF_Z(z).
\end{aligned}$$

Observing that $\int_{x=0}^{L-z} x\, dF_X(x) = (L-z)F_X(L-z) - \int_{x=0}^{L-z} F_X(x)\, dx$ and $\int_{x=L-z}^{\infty} x dF_X(x) = E[X] - \int_{x=0}^{L-z} x dF_X(x)$ we have

$$E[Y^2] = E[X^2] - 2\int_{z=0}^{\infty} \left[\left(\int_{x=L-z}^{\infty} \bar{F}_X(x) dx\right)\left(\int_{x=0}^{L-z} F_X(x)\, dx\right)\right] dF_Z(z)$$

which implies the result. □

Acknowledgments

The authors would like to thank Emre Berk for his valuable comments and suggestions during the conduct of this research. Nasuh Buyukkaramikli was partially supported by TUBITAK (The Scientific and Technological Research Council of Turkey) during the conduct of this research.

References

Alp, O., N. Erkip, R. Gullu. 2003. Optimal lot sizing/vehicle dispatching policies under stochastic lead times and stepwise fixed costs. *Oper. Res.* **51**(1): 160–166.

Alp, O., T. Tan. 2008. Tactical capacity management under capacity flexibility in make-to-stock systems. *IIE Trans.* **40**(3): 221–237.

Altiok, T. 1985. On the phase-type approximations of general distributions. *IIE Trans.* **17**(2): 110–116.

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

Axsäter, S. 1993. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Oper. Res.* **41**(4): 777–785.

Balintfy, J. L. 1964. On a basic class on inventory problems. *Manage. Sci.* **10**(2): 287–297.

Ball, M. O., B. L. Golden, A. A. Assad, L. D. Bodin. 1983. Planning for truck fleet size in the presence of a common carrier option. *Decis. Sci.* **14**(1): 103–120.

Bitran, G. R., D. Tirupati. 1988. Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Manage. Sci.* **34**(1): 75–100.

Burnson, P. 2012. 23rd Annual State of Logistics Report: Slow and Steady. Available at http://logisticsmgmt.com/images/site/LM1207_CovStateofLogistics_Rail.pdf (accessed date December 10, 2012).

Cachon, G. 2001. Managing a retailer's shelf space, inventory and transportation. *Manuf. Serv. Oper. Manag.* **3**(3): 211–229.

Cetinkaya, S., F. Mutlu, C.-Y. Lee. 2006. A comparison of outbound dispatch policies for integrated inventory and transportation decisions. *Eur. J. Oper. Res.* **171**(3): 1094–1112.

Cheung, K. L., H. Lee. 2002. The inventory benefit of shipment coordination and stock rebalancing in a supply chain. *Manage. Sci.* **48**(2): 300–306.

Dobberstein, N., C.-S. Neumann, M. Zils. 2005. Logistics in emerging markets. *McKinsey Quart.* **1**: 15–17.

Ernst, R., F. D. Pyke. 1993. Optimal base stock policies and truck capacity in a two echelon system. *Naval Res. Logist.* **40**(7): 879–903.

Federgruen, A., P. Zipkin. 1984. A combined vehicle routing and inventory allocation problem. *Oper. Res.* **32**(5): 1019–1037.

FHWA. 2005. Logistics Costs and U.S. Gross Domestic Product. Available at http://ops.fhwa.dot.gov/freight/freight_analysis/econ_methods/lcdp_rep/ (accessed date December 10, 2012).

Franx, G. J. 2001. A simple solution for the M/D/C waiting time distribution. *Oper. Res. Lett.* **29**(5): 221–229.

Graves, S. C. 1985. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Manage. Sci.* **31**(10): 1247–1256.

Gurbuz, M. C., K. Moinzadeh, Y. P. Zhou. 2007. Coordinated replenishment strategies in inventory/distribution systems. *Manage. Sci.* **53**(2): 293–307.

Ignall, E. 1969. Optimal continuous review policies for two product inventory systems with joint set-up costs. *Manage. Sci.* **15**(5): 278–283.

Mutlu, F., S. Cetinkaya S., J. H. Bookbinder. 2010. An analytical model for computing the optimal time-and-quantity-based policy for consolidated shipments. *IIE Trans.* **42**(5): 367–377.

Nielsen, C., C. Larsen. 2005. An analytical study of the q(s,s) policy applied to the joint replenishment problem. *Eur. J. Oper. Res.* **163**(3): 721–732.

Ozkaya, B. Y., U. Gurler, E. Berk. 2005. The stochastic joint replenishment problem: A new policy and analysis for single and two echelon inventory systems. Working paper, Bilkent University, Ankara, Turkey.

Ozkaya, B. Y., U. Gurler, E. Berk. 2006. The stochastic joint replenishment problem: a new policy, analysis, and insights. *Naval Res. Logist.* **53**(6): 525–546.

Pantumsinchai, P. 1992. A comparison of three joint ordering policies. *Decis. Sci.* **23**(1): 111–127.

RT. 2010. Truckers again face driver shortage, survey indicates. Refrigerated Transporter. Available at http://refrigeratedtrans.com/carriers-shippers/truckers-driver-shortage-0630/ (accessed date July 23, 2010).

Renberg, B., R. Planche. 1967. Un modle pour la gestion simultane des n articles d'un stock. *Revue Francaise d'Informatique et de Recherche Oprationelle* **6**: 47–59.

Schleyer, M., K. Furmans. 2007. An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals. *OR Spectrum.* **29**(4): 745–763.

Shang, K., S. Zhou, G.-J. van Houtum. 2010. Improving supply chain performance: Real-time demand information and flexible deliveries. *Manuf. Serv. Oper. Manag.* **12**(3): 430–448.

Sindhuchao, S., H. E. Romeijn, E. Akcali, R. Boondiskulchok. 2005. An integrated inventory-routing system for multi-item joint replenishment with limited vehicle capacity. *J. Global Optimization* **32**(1): 93–118.

Tanrikulu, M. M., A. Sen, O. Alp. 2010. A joint replenishment policy with individual control and constant size orders. *Int. J. Prod. Res.* **48**(14): 4253–4271.

Tijms, H. C. 1995. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, New York.

Toptal, A., S. Cetinkaya. 2006. Contractual agreements for coordination and vendor-managed delivery under explicit transportation considerations. *Naval Res. Logist.* **53**(5): 397–417.

Toptal, A., S. Cetinkaya, C. Y. Lee. 2003. The buyer-vendor coordination problem: Modeling inbound and outbound cargo capacity and costs. *IIE. Trans.* **35**(11): 987–1002.

Whitt, W. 1982. Approximating a point process by a renewal process, I: Two basic methods. *Oper. Res.* **30**(1): 125–147.

Whitt, W. 1993. Approximations for the GI/G/M queue. *Prod. Oper. Manag.* **2**(2): 114–161.

Yano, C. A., Y. Gerchak. 1989. Transportation contracts and safety stocks for just-in-time deliveries. *J. Manuf. Oper. Manag.* **2**(4): 314–330.