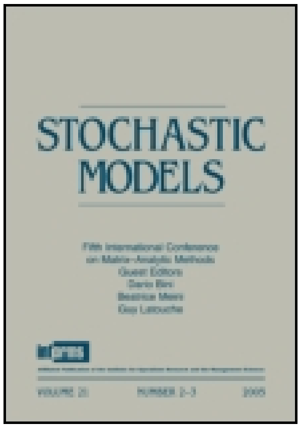

STOCHASTIC MODELS

Fifth International Conference on Matrix-Analytic Methods
Guest Editors
Dario Bini
Beatrice Meini
Guy Latouche

VOLUME 21 NUMBER 2-3 2005

Stochastic Models
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lstm20

Shipment Consolidation by Private Carrier: The Discrete Time and Discrete Quantity Case

James H. Bookbinder [a], Qishu Cai [a] & Qi-Ming He [a]

[a] Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada

Available online: 18 Nov 2011

To cite this article: James H. Bookbinder, Qishu Cai & Qi-Ming He (2011): Shipment Consolidation by Private Carrier: The Discrete Time and Discrete Quantity Case, Stochastic Models, 27:4, 664-686

To link to this article: http://dx.doi.org/10.1080/15326349.2011.614194

Stochastic Models, 27:664–686, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1080/15326349.2011.614194

SHIPMENT CONSOLIDATION BY PRIVATE CARRIER: THE DISCRETE TIME AND DISCRETE QUANTITY CASE

James H. Bookbinder, Qishu Cai, and Qi-Ming He

Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada

□ *This article studies the dispatch of consolidated shipments. Orders arrive to a depot at discrete time epochs following a discrete time batch Markov arrival process (BMAP). The weight of an order is measured in discrete units and may be correlated with the arrival time. As soon as the total weight of the accumulated orders reaches a threshold, which is a function of the time elapsed since the last dispatch, all orders are consolidated and a shipment is dispatched. A discrete time Markov chain for the accumulated weight of orders in the system is introduced and analyzed. The distributions of the accumulated weight at an arbitrary time, total accumulated weight in a consolidation cycle, and excess of weight per shipment are obtained. By introducing an absorption Markov chain and a terminating Markovian arrival process, we find the distributions of the consolidation cycle length, the waiting time of an arbitrary order, and the number of orders that occur in a cycle. An efficient computational procedure is developed for evaluating dispatch policies. The model with a quantity policy and a phase-type weight distribution is studied in detail. An extensive numerical analysis is conducted to test the efficiency of the algorithm and to gain insight into these shipment consolidation models.*

Keywords Dispatch; Freight consolidation; Markov chain; Matrix-analytic methods.

Mathematics Subject Classification 90B06; 60J10.

1. INTRODUCTION

Consider a series of small loads that arrive independently at point A, the origin, and are all destined to point B. Each load may represent an order for goods that will be assembled or picked from stock by a distributor (whose warehouse is at A). The quantity or weight of each order is a random variable, with known distribution.

Received January 2011; Accepted August 2011

Address correspondence to Qi-Ming He, Department of Management Sciences, University of Waterloo, 200 University Ave. West, Waterloo, Ontario N2L 3G1, Canada; Email: q7he@uwaterloo.ca

A *shipment consolidation* policy must decide when those orders will be sent from point A to B. Each order could be sent individually, as soon as it is ready. That would give excellent service, but at the highest transportation cost. Consolidation of orders at point A is thus preferred by the logistics firm whose own vehicle will transport those loads. The cost of transport over the fixed distance A-B is known and constant, hence the transportation cost per order would only be half as large, if even two orders were dispatched on the same truck.

More generally, we could delay the sending of a consolidated load until the total weight of those orders is a least Q; this is termed a "quantity" policy. We could instead implement a "time" policy, whereby a truck is dispatched from A to B every T periods. Because of randomness, some consolidation cycles under this policy will result in vehicles that are almost full, while other cycles will yield lighter vehicle loads. As long as the value of T is not too large, however, reasonable service can be given even to the first-arriving order, whose waiting time (before dispatch) is thus at most T.

A "hybrid" policy, often called a "time-and-quantity" policy, is also frequently employed in practice. Here a truck would be sent based on *whichever occurs first*: Either a total weight of Q is attained, or a time T has elapsed since the present consolidation cycle began. We remark that, even when the respective cost parameters are the same in each case, the optimal choices for (Q, T) in the time-and-quantity policy may differ from the optimal Q^* in the quantity policy and/or T^* for the optimal time policy.

The previous paragraph alludes to an optimization model whose objective is the minimization of total cost per unit time. In addition to transportation cost, inventory carrying cost is included. This linear term is proportional to the dollar value of the orders that are held and to the time interval that each is held. Inventory carrying cost, as a measure of customer disutility of waiting, encourages better decisions on service through its incorporation in the objective function.

In this article, we consider the case of "private carriage," transportation in one's own truck (e.g., Bookbinder and Higginson[11]). Not every manufacturing or logistics organization operates its own private fleet, but many such firms do, often employing a mixture of private carriage and "common carriage." In the latter case, the goods are moved by a public, for-hire trucking company, that charges according to the weight of each load. The freight rates for common-carrier transportation involve a quantity-discount structure. Because of that complication, and in the interests of brevity, we limit discussion in this article to transportation by private carriage. (Our approach and computational procedures, however, can be extended to the case of common carrier.)

Various methods of operations research have been applied to shipment consolidation problems over the years. Higginson and Bookbinder[5] developed a computer simulation model, studying the three consolidation policies mentioned above for a range of order-arrival rates and maximum holding times. Higginson and Bookbinder[6] employed a Markovian Decision Process. Bookbinder and Higginson[1] viewed the problem as a Stochastic Clearing System, while the treatment in Çetinkaya and Bookbinder[2] is that of Renewal Theory.

Which policy is "best"? The answer depends on the assumptions one is willing to make, and the cases to which one is able to apply the particular mathematical approach. For Poisson arrivals and orders of unit weight, Çetinkaya and Bookbinder[2] and Ülkü and Bookbinder[13] found that the quantity policy gives the best cost performance per unit time in the case of private carriage. Higginson and Bookbinder[5] treated only common carriage. They found that, for an empirically-supported Gamma distribution of order weights, the cost performance of a time policy can be better than that of a quantity policy for low order-arrival rates and fairly long holding times T. In other cases, however, Higginson and Bookbinder[5] found that the quantity policy usually gives the best cost performance, while the hybrid (time-and-quantity) policy furnishes the lowest mean delay per order.

Mütlü et al.[10] obtained *analytical* results on the optimal parameters to employ in a hybrid policy. Thus, when that policy is appropriate, and some particular assumptions hold, these renewal-theoretic results enable confidence in setting the values Q and T. Çetinkaya[3] has given a literature survey that is more thorough than is possible here. Publications cited by her and by us do have some interesting results. Our goal in the present article, however, is to furnish models that are more *general*, perhaps in several ways. For instance, the consolidation references cited above assume that orders arrive according to a Poisson process. Although our analysis here is limited to the case of discrete time, and the weight of each order is discrete, we treat the weight accumulation as a batch Markov arrival process. The weight of each arriving order is possibly correlated with its arrival time.

Matrix-analytic methods are used in this article (Neuts[11] and Latouche and Ramaswami[7]). Such methods make it possible to construct Markov chains for stochastic systems of interest and to develop efficient algorithms for computing performance measures. In this article, Markovian arrival processes are used to model the weight accumulation process through the arrivals of individual orders, for which phase-type distributions are used to model the weight. Utilizing the special QBD (quasi birth-and-death) structure of the Markov chain introduced for the system, matrix-geometric solutions are obtained for the weight accumulation process. Performance measures such as the mean cycle length, mean waiting times of orders, mean accumulated weight at an arbitrary time, and the mean level of excess can be obtained. Consequently, dispatch policies are evaluated efficiently.

The remainder of the article is organized as follows. In Section 2, the stochastic model of interest is introduced. A discrete time Markov chain is constructed in Section 3 for analyzing the weight accumulation process. Based on that discrete Markov chain, an absorption Markov chain and a terminating Markov arrival process are introduced. An efficient algorithm is developed for computing performance measures. In Section 4, the model with a quantity policy and a phase-type weight distribution is investigated in detail. Numerical examples are presented in Section 5 to test the efficiency of the algorithm and to gain insight into these shipment consolidation models. Section 6 concludes the article.

2. THE MODEL OF INTEREST

The model of interest has a private carriage and a general dispatch policy. Both the time and the weight of products are discrete. Orders arrive to the system from outside. In each period, orders are received and then a decision is made on whether or not a shipment should be dispatched. To make that decision, the total accumulated weight is calculated. If that amount exceeds a threshold, which in the general case is also a function of the time since the last dispatch, all outstanding orders are consolidated and a shipment is dispatched (i.e., all outstanding orders are cleared). Then the next cycle of accumulation and dispatch begins in the following period with zero initial weight. In the rest of this section, we give a detailed description on the weight accumulation process, the general dispatch policy, and the cost structure for the model of interest. We also briefly discuss some special cases, which are typical in practice.

2.1. The Weight Accumulation Process

Without loss of generality, we assume that at most one order can arrive in each period. Weight of orders accumulates according to a discrete time batch Markovian arrival process (*BMAP*) with a matrix-representation (D_0, D_n, $n = 1, 2, \ldots$), where D_0 and D_n are $m \times m$ nonnegative matrices, and m is a positive integer. The matrix D_n, for $n > 0$, is interpreted as the (matrix) probability that an order of n units in weight arrives. Define D as the sum of matrices $(D_0, D_1, \ldots)$. Then D is a stochastic matrix. The discrete time Markov chain associated with D is called the *underlying Markov chain* of the weight accumulation process. Denote by $I_a(t)$ the state of the underlying Markov chain at the beginning of period t. We assume that $\{I_a(t), t = 0, 1, 2, \ldots\}$ is an irreducible Markov chain. Then the matrix D is irreducible. Let $\boldsymbol{\theta}_a$ be the steady state distribution of the underlying Markov chain. Then $\boldsymbol{\theta}_a$ is the unique solution to the linear system $\boldsymbol{\theta}_a D = \boldsymbol{\theta}_a$

and $\boldsymbol{\theta}_a\mathbf{e} = 1$. Denote by λ_a the rate at which the weight accumulates, which is called the *weight arrival rate*. Then we have $\lambda_a = \boldsymbol{\theta}_a(\sum_{n=1}^{\infty} nD_n)\mathbf{e}$. Denote by λ_b the rate at which orders arrive; that is called the *order arrival rate*. Then we have $\lambda_b = \boldsymbol{\theta}_a(\sum_{n=1}^{\infty} D_n)\mathbf{e}$. (See Neuts[11] and Lucantoni[9] for more details on *BMAPs*.)

Example 2.1. Orders arrive according to a discrete time Markovian arrival process (*MAP*) with a matrix-representation (D_0, D_1). The weight of each order has a general discrete distribution $\{p_1, p_2, \ldots\}$, which is independent of the order arrival process. For this case, the weight accumulation process is a *BMAP* with matrix representation (D_0, p_nD_1, $n = 1, 2, \ldots$).

Example 2.2. Orders arrive according to a discrete time Markov arrival process with a matrix-representation (D_0, D_1). Weights of orders are independent discrete random variables with a common discrete time phase-type distribution (*PH*-distribution) ($\boldsymbol{\beta}$, S) (See Neuts[12] and Latouche and Ramaswami[7]). Here $\boldsymbol{\beta}$ is a stochastic vector (i.e., $\boldsymbol{\beta} \geq 0$ and $\boldsymbol{\beta}\mathbf{e} = 1$) and S is a substochastic matrix (i.e., all elements are nonnegative and all row sums are less than or equal to one). The distribution of the weight of each order is given by $p_n = \boldsymbol{\beta}S^{n-1}(I - S)\mathbf{e}$, $n = 1, 2, \ldots$, where I is the identity matrix.

2.2. General Dispatch Policy

A dispatch policy is given in terms of a function $f(j)$, where j is the time elapsed since the last shipment. As soon as the consolidated weight exceeds or is equal to level $f(j)$, all outstanding orders are consolidated and a shipment is dispatched. We shall call this *dispatch policy* $f(.)$, for which we make the following assumptions:

(a) $f(j) = q \geq 0$, for $j \geq j_q$, where j_q is a given positive integer. For a technical reason (see Theorem 3.1) and without loss of generality, we assume $j_q \geq 2$ and $f(j) \geq 1$ for $j = 1, 2, \ldots, j_q - 1$.
(b) $f(j)$ is non-increasing.

Assumptions (a) and (b) are intuitive and are commonly used. We also note that the popular *quantity policy*, $f(j) = Q$ for all $j \geq 1$, is a special case of the general dispatch policy.

Example 2.3. Let $f(j) = Q$ for $1 \leq j \leq j_q - 1$ and $f(j_q) = 0$. That is: after j_q units of time, a shipment is dispatched, regardless of the amount of the accumulated weight. We call $f(.)$ a *hybrid policy* (e.g., Mütlü et al.[10]). If Q is very large, the policy $f(.)$ approximates a *time policy*. If j_q is very large, the policy $f(.)$ approximates a quantity policy.

Naturally, in the "usual" time policy, the interval T between dispatches is expressed solely in terms of j, without reference to q or Q. That is the definition we employed in the Introduction. Similarly, in Section 1, we described the typical quantity policy with reference only to Q, not to j_q. (See, for example, Higginson and Bookbinder[5] and Çetinkaya[2].)

2.3. The Cost Structure

Four types of costs – dispatch, holding, order receiving, and transportation – are under consideration. Dispatch cost K_D is a fixed cost charged to each shipment. Holding cost is for the inventory or storage of products waiting for shipment. Denote by h the holding cost per unit weight per unit time. Order receiving cost is a fixed cost. Let K_S be the receiving cost per order. Finally, we denote by C_T the transportation cost per unit weight.

To evaluate a dispatch policy $f(.)$, the average total costs per unit time $C(f)$ is used. Since shipments occur in cycles, the average total costs per unit time in the steady state can be expressed as

$$C(f) = \frac{K_D + E[\mathit{Holding\ cost\ per\ cycle}]}{E[\mathit{cycle\ length}]} + K_S\lambda_b + C_T\lambda_a. \tag{2.1}$$

In the rest of the article, the analysis will thus be focused on the cycle length and the mean holding cost per unit time. In addition to $C(f)$, a number of performance measures shall also be investigated to gain insight into the consolidation and dispatch process.

3. THE MARKOV CHAIN OF INTEREST

To introduce a Markov chain for the weight accumulation process, we first define two system variables:

- Let $W(t)$ be the accumulated weight of all orders in the system at the beginning of period t (any order that arrives during period t is not included).
- Let $J(t)$ be the time elapsed since the last shipment, if the elapsed time is less than or equal to j_q; j_q+1, otherwise.

The status of the system at the beginning of period t can be represented by $(J(t), W(t), I_a(t))$. Note that $J(t) = 1$ implies that a new cycle starts at the beginning of period t (i.e., there was a shipment dispatched in the previous period). Thus, if $J(t) = 1$, we must have $W(t) = 0$. Also note that at the beginning of each period, we only see products that accumulated in previous periods. For technical reasons, if $f(j_q) = 0$ and $J(t) = j_q + 1$, we define $W(t) = 0$. In fact, $J(t)$ cannot reach the state $j_q + 1$ from other states of $J(t)$ for this case.

It is easy to see that the process $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ has the state space:

$$\begin{cases} \{1\} \times \{0\} \times \{1, 2, \ldots, m\}, \quad \text{for } J(t) = 1; \\ \{j\} \times \{0, 1, 2, \ldots, \max\{0, f(j-1) - 1\}\} \times \{1, 2, \ldots, m\}, \\ \quad \text{for } 2 \le J(t) = j \le j_q + 1. \end{cases}$$

We shall call $J(t)$ the level variable and $(W(t), I_a(t))$ the phase variable. The set of states with $J(t) = j$ shall be called *level* j. Since $\{I_a(t), t = 0, 1, 2, \ldots\}$ is a Markov chain, and because $W(t)$ depends only on the current accumulated weight and future arrivals, and $J(t)$ depends only on $W(t)$ and $I_a(t)$, it is readily seen that $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ is a Markov chain.

Theorem 3.1. *The process* $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ *is a Markov chain with transition probability matrix*

$$P_{TW} = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q + 1 \end{array} \begin{pmatrix} A_{1,1} & A_{1,2} & & & & \\ A_{2,1} & 0 & A_{2,3} & & & \\ \vdots & & \ddots & \ddots & & \\ \vdots & & & \ddots & \ddots & \\ A_{j_q,1} & & & & 0 & A_{j_q,j_q+1} \\ A_{j_q+1,1} & & & & & A_{j_q+1,j_q+1} \end{pmatrix}, \tag{3.1}$$

where

$$A_{1,1} = \overline{D}_{f(1)}, A_{1,2} = (D_0 \; D_1 \; D_2 \; \cdots \; \cdots \; D_{f(1)-1}); \quad \textit{for } 2 \le j \le j_q - 1,$$

$$A_{j,1} = \begin{array}{c} 0 \\ 1 \\ \vdots \\ f(j) - 2 \\ f(j) - 1 \\ f(j) \\ \vdots \\ f(j-1) - 1 \end{array} \begin{pmatrix} \overline{D}_{f(j)} \\ \overline{D}_{f(j)-1} \\ \vdots \\ \overline{D}_2 \\ \overline{D}_1 \\ \overline{D}_0 \\ \vdots \\ \overline{D}_0 \end{pmatrix},$$

$$A_{j,j+1} = \begin{array}{cc} & \begin{array}{ccccc} 0 & 1 & \cdots & \cdots & f(j)-1 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ \vdots \\ \vdots \\ f(j)-1 \\ f(j) \\ \vdots \\ f(j-1)-1 \end{array} & \left(\begin{array}{ccccc} D_0 & D_1 & \cdots & \cdots & D_{f(j)-1} \\ & D_0 & D_1 & \ddots & D_{f(j)-2} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & D_1 \\ & & & & D_0 \\ 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \cdots & 0 \end{array}\right) \end{array}; \tag{3.2}$$

for $j = j_q$ *and* $j = j_q + 1$, *if* $f(j_q) > 0$, *the matrices* $A_{j_q,1}$, A_{j_q,j_q+1}, $A_{j_q+1,1}$, *and* A_{j_q+1,j_q+1} *are given by equation* (3.2); *if* $f(j_q) = 0$, *then* $A_{j_q,1}$ *is given by equation* (3.2), $A_{j_q,j_q+1} = 0$, $A_{j_q+1,1} = \overline{D}_0$, *and* $A_{j_q+1,j_q+1} = 0$. *Note that* $\overline{D}_n = \sum_{j=n}^{\infty} D_j$, $n = 0, 1, \ldots$.

Proof. The transitions between levels $j = 1, 2, \ldots, j_q$, can be identified based on the following observations:

i) The value of $J(t)$ always increases by 1, except for possible transitions to level 1 when a shipment is dispatched.
ii) The value of $W(t)$ is non-decreasing, except for possible transitions to level 1 when a shipment is dispatched.
iii) If $J(t) = j$, the initial weight in period t is between 0 and $f(j-1)-1$, and the ending weight is between 0 and $f(j)-1$.

The transitions associated with level $j_q + 1$ are based on the fact that for the policy $f(j)$, the dispatch quantity is the same for $j \geq j_q$. If $f(j_q) = 0$, a shipment must be dispatched once the time elapsed since the last shipment is j_q. Thus, there is no transition from level j_q to level $j_q + 1$. The transition probabilities are obtained accordingly. □

For the case of $f(j_q) = 0$, states in level $j_q + 1$ are overflow states. Those states do not affect the analysis, since transition probabilities from level j_q to level $j_q + 1$ are zero. In the following analysis, for convenience, we assume that the Markov chain $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ has $j_q + 1$ levels.

Denote by $\boldsymbol{\theta}_{TW}$ the steady state distribution of $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$. Then $\boldsymbol{\theta}_{TW}$ is the unique solution to the linear system $\boldsymbol{\theta}_{TW} P_{TW} = \boldsymbol{\theta}_{TW}$ and $\boldsymbol{\theta}_{TW}\mathbf{e} = 1$. We decompose $\boldsymbol{\theta}_{TW}$ as follows: $\boldsymbol{\theta}_{TW} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{j_q}, \boldsymbol{\theta}_{j_q+1})$, and $\boldsymbol{\theta}_j = (\boldsymbol{\theta}_{j,0}, \boldsymbol{\theta}_{j,1}, \ldots, \boldsymbol{\theta}_{j,\max\{0,f(j-1)-2\}}, \boldsymbol{\theta}_{j,\max\{0,f(j-1)-1\}})$, for $j = 2, 3, \ldots, j_q + 1$. We note that all vectors $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_{j,w}, w = 0, 1, \ldots, \max\{0, f(j-1)-1\}$, and $j = 2, 3, \ldots, j_q + 1\}$ are row vectors of size m_a. Because of the special structure in P_{TW}, an efficient algorithm can be developed for computing $\boldsymbol{\theta}_{TW}$, if $\boldsymbol{\theta}_{TW}$ exists.

Algorithm I.

(I.1) Compute the matrices, $R_1 = I$,

$$\begin{aligned} R_j &= R_{j-1}A_{j-1,j}, \quad 2 \le j \le j_q; \\ R_{j_q+1} &= R_{j_q}A_{j_q,j_q+1}(I - A_{j_q+1,j_q+1})^{-1}; \\ P_1 &= \sum_{j=1}^{j_q+1} R_j A_{j,1}. \end{aligned} \tag{3.3}$$

(I.2) Solve the linear system $\boldsymbol{\theta}_1 P_1 = \boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_1\left(\sum_{j=1}^{j_q+1} R_j\mathbf{e}\right) = 1$ for $\boldsymbol{\theta}_1$.
(I.3) Compute $\boldsymbol{\theta}_j = \boldsymbol{\theta}_1 R_j$, for $j = 1, 2, \ldots, j_q + 1$.

The validity of Algorithm I is guaranteed by the finiteness of the number of states and the fact that states in level 1 can be reached from other states.

Theorem 3.2. *The steady state distribution of the Markov chain* $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ *exists and is given by* $\boldsymbol{\theta}_{TW}$.

Proof. First, it is easy to verify that $\boldsymbol{\theta}_{TW}$ is a solution to the linear system $\boldsymbol{\theta}_{TW}P_{TW} = \boldsymbol{\theta}_{TW}$ and $\boldsymbol{\theta}_{TW}\mathbf{e} = 1$. Since the underlying Markov chain $\{I_a(t), t = 0, 1, 2, \ldots\}$ is irreducible, the Markov chain $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ has a single closed set. Consequently, limiting probabilities exist and are unique, and hence form the steady state distribution. This also implies that the solution to the linear system is unique. Thus, $\boldsymbol{\theta}_{TW}$ is the steady state distribution. □

As an immediate consequence of Theorem 3.2, $\boldsymbol{\theta}_a = \sum_{j=1}^{j_q+1}\sum_{w=0}^{\max\{0,f(j-1)-1\}} \boldsymbol{\theta}_{j,w}$ (note: $f(0) = 1$), which can be used for checking accuracy in computation. A number of performance measures can be found directly or indirectly from $\boldsymbol{\theta}_{TW}$.

Excess O_w is defined as the amount of weight over a threshold function $g(j)$ at a dispatch epoch. Typical choices of the function $g(j)$ include $g(.) = f(.)$ or a constant function $g(j) = Q_o$, for all j. If the constant Q_o were the truck size, an excess would lead to extra transportation costs.

Denote by W the accumulated weight at the beginning of an arbitrary period. Let W_c be the accumulated weight of an arbitrary shipment (i.e., the total weight accumulated during an arbitrary cycle). The distributions of W, O_w, and W_c can be obtained from $\boldsymbol{\theta}_{TW}$ in a straightforward manner.

Denote by P_S the probability that a shipment takes place in an arbitrary time period. The following results are obtained from Theorems 3.1 and 3.2.

Corollary 3.3. *For the shipment consolidation model defined in Section 2, we have*

(i)

$$P\{W = w\} = \left(\sum_{j=1:w \leq f(j-1)-1}^{j_q+1} \boldsymbol{\theta}_{j,w} \right) \mathbf{e}, \quad w = 0, 1, 2, \ldots, f(1) - 1.$$

(ii)

$$P_S = \boldsymbol{\theta}_1 \mathbf{e}.$$

(iii)

$$P\{W_c = i\} = \frac{1}{P_S} \sum_{j=1}^{j_q+1} \left(\sum_{w=0:i \geq f(j) \ and \ i \geq w}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} D_{i-w} \mathbf{e} \right), \quad i = 0, 1, 2, \ldots.$$

(iv)

$$P\{O_w = i\} = \frac{1}{P_S} \sum_{j=1}^{j_q+1} \left(\sum_{w=0:i+g(j) \geq f(j)}^{\min\{i+g(j), f(j-1)-1\}} \boldsymbol{\theta}_{j,w} D_{i+g(j)-w} \mathbf{e} \right),$$

$$i = 0, 1, 2, \ldots; \quad and$$

$$P_o = P\{O_w > 0\} = \frac{1}{P_S} \sum_{j=1}^{j_q+1} \left(\sum_{w=0}^{f(j-1)-1} \boldsymbol{\theta}_{j,w} \overline{D}_{\max\{0, \max\{1+g(j), f(j)\}-w\}} \mathbf{e} \right).$$

The means $E[W]$, $E[W_c]$, and $E[O_w]$ can be obtained from the distributions accordingly.

Remark 3.1. By definition, immediately after a shipment takes place, the clock for a new consolidation cycle is set to one (i.e., $J(t) = 1$). Thus, the probability P_S that a shipment takes place is equal to the probability that the elapsed time since the last shipment is one (i.e., $\boldsymbol{\theta}_1 \mathbf{e}$). This gives an intuitive interpretation to part (ii) of Corollary 3.3.

For a number of special cases, the computation of $\boldsymbol{\theta}_{TW}$ and performance measures can be simplified significantly.

Case 1. Quantity Policy Model Consider a model with a quantity policy $f(.)$, i.e., $f(j) = Q$ for all $j \geq 1$. For this special case, we can set $j_q = 1$. Then the Markov chain $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ need have only two levels: $J(t) = 1$ and 2. In addition, the size of matrices involved in computation is m, instead of Qm required in Algorithm I.

Case 2. Hybrid Policy Model (Example 2.3) Consider a model with a hybrid policy $f(.)$, i.e., $f(j) = Q$ for $1 \leq j \leq j_q - 1$ and $f(j) = 0$ for $j \geq j_q \geq 2$. For this case, the Markov chain $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ need have only j_q levels, instead of $j_q + 1$. All levels, except for level 1, have the same number of phases.

Case 3. Independent-Weights Model (Example 2.1) Consider the weight arrival process defined in Example 2.1, where the order weights are independent of the order arrival process. Then we have, $\overline{D}_0 = D_0 + D_1$ and $\overline{D}_n = (1 - p_1 - p_2 - \cdots - p_{n-1})D_1$, $n = 1, 2, \ldots$. (Note $p_0 = 0$.) Consequently, construction of the transition probability matrix and all expressions given in Corollary 3.3 can be reduced to finite summations. Furthermore, the expressions of $E[W]$, $E[W_c]$, and $E[O_w]$ can be simplified to finite summations.

Next, we construct an *absorption Markov chain* to investigate the length of a consolidation cycle L_c and the waiting time L_w of an arbitrary order. Define

$$T_c = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \begin{pmatrix} 0 & A_{1,2} & & & & \\ & 0 & A_{2,3} & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 0 & A_{j_q,j_q+1} \\ & & & & & A_{j_q+1,j_q+1} \end{pmatrix}, \quad T_c^0 = \begin{array}{c} 1 \\ 2 \\ \vdots \\ \vdots \\ j_q \\ j_q+1 \end{array} \begin{pmatrix} A_{1,1} \\ A_{2,1} \\ \vdots \\ \vdots \\ A_{j_q,1} \\ A_{j_q+1,1} \end{pmatrix}, \qquad (3.4)$$

and $P_c = (I, 0, \ldots, 0)(I - T_c)^{-1}T_c^0$. The matrix P_c is a stochastic matrix that governs the transitions of the underlying Markov chain $\{I_a(t), t = 0, 1, 2, \ldots\}$ at the beginnings of consolidation cycles (or at the ends of those cycles), i.e., the embedded Markov chain for the state of the underlying Markov chain of the weight arrival process at the beginnings of consolidation cycles. It is readily seen that $P_{TW} = T_c + (T_c^0, 0, \ldots, 0)$. Denote by $\boldsymbol{\eta}_c$ the steady state distribution associated with P_c. Then $\boldsymbol{\eta}_c$ is the unique solution to the linear system $\boldsymbol{\eta}_c P_c = \boldsymbol{\eta}_c$ and $\boldsymbol{\eta}_c \mathbf{e} = 1$.

Theorem 3.4. *The distribution $\boldsymbol{\eta}_c$ is given by $\boldsymbol{\eta}_c = \boldsymbol{\theta}_1/(\boldsymbol{\theta}_1\mathbf{e})$. In steady state, the distribution of the length of a consolidation cycle L_c has a discrete time phase-type distribution with matrix representation $((\boldsymbol{\eta}_c, 0, \ldots, 0), T_c)$. The distribution of L_w+1* *has a phase-type distribution with matrix representation* $(\boldsymbol{\theta}_{TW}, T_c)$. *In addition, we have*

$$E[L_c] = 1/(\boldsymbol{\theta}_1\mathbf{e});$$
$$E[L_w] = \sum_{j=1}^{j_q} j\boldsymbol{\theta}_j\mathbf{e} + j_q\boldsymbol{\theta}_{j_q+1}\mathbf{e} + \boldsymbol{\theta}_{j_q+1}(I - A_{j_q+1,j_q+1})^{-1}\mathbf{e} - 1. \tag{3.5}$$

Proof. Due to the special structure within the matrix T_c, the first row of the inverse of $I - T_c$ can be found explicitly as $(R_1, R_2, \ldots, R_{j_q+1})$. Immediately, we obtain $P_c = \sum_{j=1}^{j_q+1} R_j A_{j,1}$. The vector $\boldsymbol{\eta}_c$ is the unique solution to the linear system $\boldsymbol{\eta}_c\left(\sum_{j=1}^{j_q+1} R_j A_{j,1}\right) = \boldsymbol{\eta}_c$ and $\boldsymbol{\eta}_c\mathbf{e} = 1$. Existence of the steady state distribution is again guaranteed by the fact that P_c has a single closed set of states. By Algorithm I, it is easy to see that $\boldsymbol{\eta}_c = \boldsymbol{\theta}_1/(\boldsymbol{\theta}_1\mathbf{e})$. The initial probability distribution of the Markov chain $\{(J(t), W(t), I_a(t)), t = 0, 1, 2, \ldots\}$ at the beginning of a consolidation cycle is given by $(\boldsymbol{\eta}_c, 0, \ldots, 0)$, since $J(t) = 1$ at the beginning of any cycle. Thus, the length of a consolidation cycle has a phase-type distribution. The average cycle length is obtained by straightforward simplification of the expression $(\eta_c, 0, \ldots, 0)(I - T_c)^{-1}\mathbf{e}$, which leads to $E[L_c] = \eta_c\left(\sum_{j=1}^{j_q+1} R_j\mathbf{e}\right) = (\boldsymbol{\theta}_1\mathbf{e} + \boldsymbol{\theta}_2\mathbf{e} + \cdots + \boldsymbol{\theta}_{j_q+1}\mathbf{e})/(\boldsymbol{\theta}_1\mathbf{e}) = 1/(\boldsymbol{\theta}_1\mathbf{e})$.

For L_w, note that the waiting time is zero if an order arrives and a shipment is dispatched in the same period. By some routine calculations, the results can be obtained. □

Remark 3.2.

i) That $E[L_c] = 1/(\boldsymbol{\theta}_1\mathbf{e})$ can be explained intuitively. The probability that the system is at the beginning of a consolidation cycle is $\boldsymbol{\theta}_1\mathbf{e}$. Thus, the mean time between consecutive visits to such states is given by $1/(\boldsymbol{\theta}_1\mathbf{e})$.
ii) From the mean total consolidated weight per cycle $E[W_c]$ (Corollary 3.3) and the mean cycle length $E[L_c]$ (Theorem 3.4), the mean weight shipped per unit time can be obtained as $E[W_c]/E[L_c]$. As indicated in Section 2, the mean weight that arrives per unit time is given by λ_a. Then we must have $\lambda_a = E[W_c]/E[L_c]$. Such a relationship is useful for checking the accuracy of numerical computations.

Finally in this section, we introduce a *terminating Markov arrival process* (He and Neuts (1998)) to study the number of orders N_c received in a consolidation cycle. First, we decompose matrices T_c and T_c^0 defined in equation (3.4) as follows:

$$T_{c,0} = \begin{array}{c} 1\\ 2\\ \vdots\\ \vdots\\ j_q\\ j_q+1 \end{array} \begin{pmatrix} 0 & A_{1,2,0} & & & & \\ & 0 & A_{2,3,0} & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 0 & A_{j_q,j_q+1,0} \\ & & & & & A_{j_q+1,j_q+1,0} \end{pmatrix}, \quad T_{c,0}^0 = \begin{array}{c} 1\\ 2\\ \vdots\\ \vdots\\ j_q\\ j_q+1 \end{array} \begin{pmatrix} A_{1,1,0} \\ A_{2,1,0} \\ \vdots \\ \vdots \\ A_{j_q,1,0} \\ A_{j_q+1,1,0} \end{pmatrix}; \tag{3.6}$$

$$T_{c,1} = \begin{array}{c} 1\\ 2\\ \vdots\\ \vdots\\ j_q\\ j_q+1 \end{array} \begin{pmatrix} 0 & A_{1,2,1} & & & & \\ & 0 & A_{2,3,1} & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 0 & A_{j_q,j_q+1,1} \\ & & & & & A_{j_q+1,j_q+1,1} \end{pmatrix}, \quad T_{c,1}^0 = \begin{array}{c} 1\\ 2\\ \vdots\\ \vdots\\ j_q\\ j_q+1 \end{array} \begin{pmatrix} A_{1,1,1} \\ A_{2,1,1} \\ \vdots \\ \vdots \\ A_{j_q,1,1} \\ A_{j_q+1,1,1} \end{pmatrix},$$

where $A_{j,j+1,0}$ and $A_{j,1,0}$ are respectively obtained from $A_{j,j+1}$ and $A_{j,1}$ by keeping only block D_0, while $A_{j,j+1,1}$ and $A_{j,1,1}$ are obtained respectively from $A_{j,j+1}$ and $A_{j,1}$ by removing the block D_0. By definition, we must have $A_{j,j+1} = A_{j,j+1,0} + A_{j,j+1,1}$, $A_{j,1} = A_{j,1,0} + A_{j,1,1}$, $T_c = T_{c,0} + T_{c,1}$, and $T_c^0 = T_{c,0}^0 + T_{c,1}^0$. We consider a terminating Markov arrival process defined by $(T_{c,0}, T_{c,1}, T_{c,0}^0, T_{c,1}^0)$, where $T_{c,1}$ and $T_{c,1}^0$ correspond to transitions with order arrivals, and $T_{c,0}$ and $T_{c,0}^0$ correspond to transitions without order arrivals. This Markov arrival process is called a *terminating process* since we only count the number of order arrivals before or at the time the process enters level one.

Theorem 3.5. *Given an initial probability distribution* $((\boldsymbol{\theta}_1/(\boldsymbol{\theta}_1\mathbf{e}), 0, \ldots, 0)$*, the number of orders that occur in a consolidation cycle* N_w *equals the total number of arrivals in the terminating Markov arrival process* $(T_{c,0}, T_{c,1}, T_{c,0}^0, T_{c,1}^0)$. *Consequently, in steady state, we have*

$$P\{N_c = n\} = \begin{cases} \left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathbf{e}}, 0, \ldots, 0\right)(I - T_{c,0})^{-1}T_{c,0}^0\mathbf{e}, & n = 0; \\[2ex] \left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathbf{e}}, 0, \ldots, 0\right)((I - T_{c,0})^{-1}T_{c,1})^n(I - T_{c,0})^{-1}T_{c,0}^0\mathbf{e} & \\[2ex] + \left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathbf{e}}, 0, \ldots, 0\right)((I - T_{c,0})^{-1}T_{c,1})^{n-1}(I - T_{c,0})^{-1}T_{c,1}^0\mathbf{e}, & n \geq 1. \end{cases} \tag{3.7}$$

The mean number of orders per cycle is given by

$$E[N_c] = \frac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathbf{e}}\left(\sum_{j=1}^{j_q} R_j A_{j,j+1,1}\mathbf{e}+R_{j_q+1}A_{j_q+1,j_q+1,1}\mathbf{e}\right)+\frac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_1\mathbf{e}}\left(\sum_{j=1}^{j_q+1} R_j A_{j,1,1}\mathbf{e}\right). \tag{3.8}$$

Proof. The terminating process $(T_{c,0},\ T_{c,1},\ T^0_{c,0},\ T^0_{c,1})$ is a special discrete version of the terminating Markov arrival process defined in He and Neuts (Ref.[4]) (See also Latouche et al. (Ref.[8])). The distribution and the mean of N_c are obtained routinely. □

Remark 3.3.

i) The sum of the numerators in equation (3.8) is the probability that an order arrives in an arbitrary time period, which is also the expected number of arrivals in that period. Multiplying that sum by the mean cycle length yields the total number of order arrivals in an arbitrary cycle.
ii) By Theorems 3.4 and 3.5, the number of orders per unit time is given by $E[N_c]/E[L_c]$. Then we must have $\lambda_b = E[N_c]/E[L_c]$.

4. QUANTITY POLICY AND PHASE-TYPE WEIGHT MODEL

Consider a quantity policy model $f(j) = Q$ for all $j \geq 1$. The weight arrival process is given by $(D_0,\ \boldsymbol{\beta}S^{n-1}(I - S)\mathbf{e}D_1,\ n = 1, 2, \ldots)$ defined in Examples 2.1 and 2.2. We use an alternative approach to analyze this case. We begin by showing an explicit result for the excess O_w.

Theorem 4.1. *Assume that the threshold function for excess is $g(j) = Q_o$ for all j. If $Q_o \geq Q$, the excess at a dispatch epoch has a phase-type distribution with matrix representation $(\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}, S)$, where* $\mathbf{S}^0 = \mathbf{e} - S\mathbf{e}$. *In addition, we have $E[O_w] = \boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}(I - S)^{-1}\mathbf{e}$, and $P\{O_w > 0\} = \boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}\mathbf{e}$. If $Q_o < Q$, then $O_w = Q - Q_o + O_{w,Q}$, where $O_{w,Q}$ is the excess if $Q_o = Q$.*

Proof. Consider a Markov arrival process with matrix presentation $(S, \mathbf{S}^0\boldsymbol{\beta})$ (i.e., a *PH*-renewal process (see Neuts Ref.[12])). If we treat Q_o as time, then the excess at Q_o is the time until the next arrival of that Markovian arrival process. If $Q_o \geq Q$, the distribution of the phase at Q_o is $\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}$. Then the excess at Q_o has a discrete time phase-type distribution with matrix representation $(\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S^{Q_o-Q+1}, S)$.

The mean excess and the probability that an excess occurs are obtained accordingly. Results for the case $Q_o < Q$ are obtained by noticing that a dispatch occurs if and only if the accumulated weight is greater than or equal to Q. □

By taking advantage of the partial memoryless property of the phase-type distributions, a new discrete time Markov chain can be introduced for the weight process. The idea is: After every order arrival, stop the clock of the order arrival process and start a fictitious clock for the underlying discrete time Markov chain for the weight distribution. The fictitious clock is stopped, and the clock of the order arrival process resumes, when the underlying Markov chain of the phase-type distribution reaches its absorption state.

More specifically, let $\{I_w(t), t = 0, 1, 2, \ldots\}$ be the phase of the underlying Markov chain for the phase-type distribution $(\boldsymbol{\beta}, S)$ before absorption. Then a new Markov chain can be constructed in the following way.

1) If an order arrives, the underlying Markov chain $\{I_w(t), t = 0, 1, 2, \ldots\}$ is turned on, initialized by $\boldsymbol{\beta}$, immediately and the underlying Markov chain $\{I_a(t), t = 0, 1, 2, \ldots\}$ is frozen.
2) If the underlying Markov chain $\{I_w(t), t = 0, 1, 2, \ldots\}$ enters its absorption state, it is terminated, and the Markov chain $\{I_a(t), t = 0, 1, 2, \ldots\}$ is unfrozen.

Define

$\widehat{I}_a(t)$: $\widehat{I}_a(t) = I_a(t)$, if the clock of the order arrival process is on; otherwise, $\widehat{I}_a(t)$ is the last value of $I_a(t)$ before $I_a(t)$ is frozen.
$\widehat{I}_w(t)$: $\widehat{I}_w(t) = I_w(t)$, if the clock of the phase-type distribution is on; but $\widehat{I}_w(t) = 0$, if the clock of the order arrival process is on.
$\widehat{W}(t)$: $\widehat{W}(t) = W(t)$, if the clock of the order arrival process is on; otherwise, if the clock of the phase-type distribution is on, $\widehat{W}(t)$ increases by one per unit time if $\widehat{W}(t-1) < Q - 1$, and becomes 0 if $\widehat{W}(t-1) = Q - 1$.

Note that $W(t)$ takes values $\{0, 1, 2, \ldots, Q-1\}$. Then the process $\{(\widehat{W}(t), \widehat{I}_a(t), \widehat{I}_w(t)), t = 0, 1, 2, \ldots\}$ is a Markov chain with transition probability matrix $P_{TW} = D$ for Q =1, and, for $Q \geq 2$,

$$P_{TW} = \begin{array}{r} (0, i_a) \\ (1, i_a) \\ (1, i_a, i_w) \\ (2, i_a) \\ (2, i_a, i_w) \\ \vdots \\ (Q-2, i_a) \\ (Q-2, i_a, i_w) \\ (Q-1, i_a) \\ (Q-1, i_a, i_w) \end{array} \begin{pmatrix} D_0 & (0, D_1 \otimes \boldsymbol{\beta}) & & & & \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & D_1 \otimes \boldsymbol{\beta} \\ 0 & I \otimes S \end{pmatrix} & & & \\ \vdots & & \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & D_1 \otimes \boldsymbol{\beta} \\ 0 & I \otimes S \end{pmatrix} & & \\ \vdots & & & \ddots & \ddots & \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & & & & \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & D_1 \otimes \boldsymbol{\beta} \\ 0 & I \otimes S \end{pmatrix} \\ \begin{pmatrix} D_1 \\ I \otimes (S\mathbf{e}) \end{pmatrix} & & & & & \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix} \end{pmatrix} \tag{4.1}$$

Let $\boldsymbol{\pi}_{TW}$ be the steady state distribution of P_{TW}, i.e., $\boldsymbol{\pi}_{TW} P_{TW} = \boldsymbol{\pi}_{TW}$ and $\boldsymbol{\pi}_{TW}\mathbf{e} = 1$. We decompose $\boldsymbol{\pi}_{TW}$ as follows: $\boldsymbol{\pi}_{TW} = (\boldsymbol{\pi}_0, (\boldsymbol{\pi}_{1,a}, \boldsymbol{\pi}_{1,w}), (\boldsymbol{\pi}_{2,a}, \boldsymbol{\pi}_{2,w}), \ldots, (\boldsymbol{\pi}_{Q-2,a}, \boldsymbol{\pi}_{Q-2,w}), (\boldsymbol{\pi}_{Q-1,a}, \boldsymbol{\pi}_{Q-1,w}))$. The steady state distribution $\boldsymbol{\pi}_{TW}$ can be computed by using the following algorithm.

Algorithm I(PH).

(I(PH).1) Compute the matrices $X_0 = I$,

$$\begin{aligned} X_1 &= X_0 \left(0 \; D_1 \otimes \boldsymbol{\beta}\right) \left(I - \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix}\right)^{-1}, \\ X_w &= X_{w-1} \begin{pmatrix} 0 & D_1 \otimes \boldsymbol{\beta} \\ 0 & I \otimes S \end{pmatrix} \left(I - \begin{pmatrix} D_0 & 0 \\ I \otimes \mathbf{S}^0 & 0 \end{pmatrix}\right)^{-1}, \quad 2 \leq w \leq Q-1; \\ P_1 &= D_0 + X_{Q-1} \begin{pmatrix} D_1 \\ I \otimes S\mathbf{e} \end{pmatrix}. \end{aligned} \tag{4.2}$$

(I(PH).2) Solve the linear system $\boldsymbol{\pi}_0 P_1 = \boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_0(\sum_{j=0}^{Q-1} X_j)\mathbf{e} = 1$ for $\boldsymbol{\pi}_0$.
(I(PH).3) Compute $(\boldsymbol{\pi}_{i,a}, \boldsymbol{\pi}_{i,w}) = \boldsymbol{\pi}_0 X_i$, for $i = 1, 2, \ldots, Q-1$.

Compared to Algorithm I, Algorithm I(PH) computes matrices only of the size m or mm_b, where m_b is the size of S. Similar to Theorem 3.4, the mean cycle length between shipments can be found as follows.

Theorem 4.2. *The mean time between two consecutive entrances to level zero from level $Q-1$ of the Markov chain $\{(\widehat{W}(t), \widehat{I}_a(t), \widehat{I}_w(t)), t = 0, 1, 2, \ldots\}$ is given by $1/(\boldsymbol{\pi}_0 D_1 \mathbf{e})$. Consequently, the mean cycle length of consolidated shipments is $E[L_c] = 1/(\boldsymbol{\pi}_0 D_1 \mathbf{e}) + 1 - Q$.*

Proof. Similar to the proof Theorem 3.4 and by the structure of P_{TW}, the transition probability matrix of the embedded Markov chain at the end of each consolidation cycle is given by $P_2 = (I - D_0)^{-1} X_{Q-1}\binom{D_1}{I \otimes S\mathbf{e}}$. Let $\boldsymbol{\eta}_0$ be the invariant vector of P_2, i.e., $\boldsymbol{\eta}_0 P_2 = \boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_0 \mathbf{e} = 1$. By equation (4.10), it can be shown that $\boldsymbol{\eta}_0(I - D_0)^{-1} = \delta\boldsymbol{\pi}_0$, which leads to $\boldsymbol{\eta}_0 = \boldsymbol{\pi}_0(I - D_0)/(\boldsymbol{\pi}_0(I - D_0)\mathbf{e})$ and $\delta = 1/(\boldsymbol{\pi}_0(I - D_0)\mathbf{e})$. Similar to Theorem 3.4, the mean time between two consecutive entrances from level $Q - 1$ to level zero can be obtained as $1/(\boldsymbol{\pi}_0(I - D_0)\mathbf{e})$. Intuitively, the probability that the Markov chain just entered level zero is $\boldsymbol{\pi}_0(I - D_0)\mathbf{e}$. Thus, the average time between two consecutive entrances to level zero from level $Q - 1$ is $1/(\boldsymbol{\pi}_0(I - D_0)\mathbf{e})$. Since the fictitious time between two consecutive visits to level zero from level $Q - 1$ is exactly $Q - 1$ (i.e., the accumulated weight increases to Q), the average cycle length of shipments is obtained as $E[L_c] = 1/(\boldsymbol{\pi}_0(I - D_0)\mathbf{e}) - (Q - 1)$, which, together with $(D_0 + D_1)\mathbf{e} = \mathbf{e}$, leads to the expected result. □

The consolidated weight per cycle equals Q plus possible overshot above the level Q. By Theorem 4.1, the consolidated weight per cycle is expressed as Q plus a phase-type variable with matrix representation $(\boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S, S)$. The mean consolidated weight per cycle is given by

$$E[W_c] = Q + \boldsymbol{\beta}(S + \mathbf{S}^0\boldsymbol{\beta})^{Q-1}S(I - S)^{-1}\mathbf{e}. \tag{4.3}$$

To find the accumulated weight at an arbitrary time, we consider the steady state distribution for the process $\{(W(t), I_a(t)), t = 0, 1, 2, \dots\}$ (i.e., censoring out all the phases associated with the underlying Markov chain $\{I_w(t), t = 0, 1, 2, \dots\}$). Define

$$\begin{aligned}\boldsymbol{\pi}_{TW,a} &= (\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_{Q-2}, \boldsymbol{\phi}_{Q-1})\\ &= (\boldsymbol{\pi}_0, \boldsymbol{\pi}_{1,a}, \boldsymbol{\pi}_{2,a}, \dots, \boldsymbol{\pi}_{Q-2,a}, \boldsymbol{\pi}_{Q-1,a})/((\boldsymbol{\pi}_0 + \boldsymbol{\pi}_{1,a} + \boldsymbol{\pi}_{2,a}\\ &\qquad + \cdots + \boldsymbol{\pi}_{Q-2,a} + \boldsymbol{\pi}_{Q-1,a})\mathbf{e}).\end{aligned} \tag{4.4}$$

By definition, $\boldsymbol{\pi}_{TW,a}$ is the steady state distribution of the total weight $W(t)$ at an arbitrary time. Then we obtain $E[W] = \sum_{i=0}^{Q-1} i\boldsymbol{\phi}_i\mathbf{e}$. In addition, we must have $\boldsymbol{\theta}_a = \boldsymbol{\phi}_0 + \boldsymbol{\phi}_1 + \boldsymbol{\phi}_2 + \cdots + \boldsymbol{\phi}_{Q-2} + \boldsymbol{\phi}_{Q-1}$, which is useful for checking computational accuracy.

5. NUMERICAL ANALYSIS

By the definition given in Section 2, the average total costs per unit time for a dispatch policy $f(.)$ can be obtained by using equation (2.1) with $E[\textit{cycle length}] = E[L_c]$ and $E[\textit{holding cost per cycle}] = hE[W]E[L_c]$. Based on results given in Corollary 3.3 and Theorem 3.4, the steps for computing $C(f)$ can be summarized as follows.

5.1. Computational Procedure

Parameters: $(D_0, D_n, n = 1, 2, \ldots), f(.), g(.), K_D, h, K_S$, and C_T.

1) Determine $\boldsymbol{\theta}_a$, λ_a, and λ_b;
2) Use Algorithm I to compute the vector $\boldsymbol{\theta}_1$ and $(R_1, R_2, \ldots, R_{j_q+1})$;
3) Use Corollary 3.3 to calculate $E[W]$, $P\{O_w > 0\}$, and $E[O_w]$;
4) Use Theorems 3.4 and 3.5 to obtain $E[L_c]$, $E[L_w]$, and $E[N_c]$; and finally,
5) Use equation (2.1) to compute $C(f)$.

The weight arrival processes for Examples 5.1 and 5.2 are given as follows. The selected arrival processes are quite different in terms of the correlation between arrival times and batch sizes and the variance of batch sizes. We discuss how such characteristics affect the performance of dispatch policies.

5.2. Correlated Arrival Process (*CAP)*

In this process, the arrival times and the weights of orders are correlated. The matrix representation of the *BMAP* is

$$D_0 = \begin{pmatrix} 0 & 0.3 & 0 & 0 & 0 \\ 0 & 0.1 & 0.6 & 0 & 0 \\ 0 & 0 & 0.1 & 0.6 & 0 \\ 0 & 0 & 0 & 0.1 & 0.7 \\ 0.8 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad D_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad (5.1)$$

$$D_4 = \begin{pmatrix} 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad D_5 = \begin{pmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The order arrival rate is $\lambda_b = 0.3354$ and the weight arrival rate is $\lambda_a = 1.0354$. Note that large orders occur only if the phase of the underlying Markov chain is either 1 or 2.

5.3. Heavy Tailed Arrival Process (*HTAP)*

In this process, the order arrival times and the order weights are independent. The weight distribution has a heavy tail. Orders arrive according to an *MAP* with matrix representation

$$D_0 = \begin{pmatrix} 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0.1 & 0.2 & 0 & 0 \\ 0 & 0 & 0.1 & 0.4 & 0 \\ 0 & 0 & 0 & 0.1 & 0.5 \\ 0.6 & 0 & 0 & 0 & 0.1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \\ 0.3 & 0 & 0 & 0 & 0 \end{pmatrix}. \quad (5.2)$$

Let ω be a random variable whose distribution is the normalized Riemann Zeta function $\{p_n = 1/(\zeta_{2.5} n^{2.5}),\ n = 1, 2, \ldots\}$, where $\zeta_{2.5} = 1.3415$ is the normalization factor. The weight w_o of an order is defined as $w_o = 2\omega$. The order arrival rate is $\lambda_b = 0.5347$, the average weight per order is $E[w_o] = 1.9325$, and the weight arrival rate is $\lambda_a = \lambda_b E[w_o] = 1.0333$. The standard deviation of w_o is 12.01.

5.4. Light Tailed Arrival Process *LTAP*

This arrival process is similar to the HTAP, except that the weight of an order has a phase-type distribution with matrix representation $(\boldsymbol{\beta}, S)$, where $\boldsymbol{\beta} = (0.1\ \ 0.9), S = \left(\begin{smallmatrix} 0.15 & 0.3 \\ 0.2 & 0.3 \end{smallmatrix}\right)$. It is well-known that phase-type distributions are light tailed. The order arrival rate is $\lambda_b = 0.5347$, the mean weight per order is $E[w_o] = 1.9533$, and the weight arrival rate is $\lambda_a = 1.0444$. Here, the standard deviation of w_o is 1.3458, which is significantly smaller than that in the case of HTAP.

Example 5.1. We consider models where the weight arrival processes are *CAP*, *HTAP*, and *LTAP*, and a quantity policy with threshold Q. We assume that the excess threshold Q_o is fixed at 20 for all cases. Cost parameters are: $K_D = 10$ and $h = 0.1$. Since the order-receiving cost and transportation cost are approximately the same for all models and all policies, without loss of generality, we assume $K_S = 0$ and $C_T = 0$.

Applying Algorithm I, the average total costs per unit time are computed for $Q = 1, 2, \ldots, 40$, and the results are plotted in Figure 1.

Figure 1 demonstrates that, as a function of Q, $C(f)$ is unimodal, but may not be convex. The optimal Qs, for which $C(f)$ is minimized, are 13, 12, and 14 for models with CAP, HTAP, and LTAP, respectively. It is interesting to see that the HTAP case has the lowest cost for many Qs. By looking into the individual costs per unit time, it seems that the HTAP case has a smaller holding cost per unit time. The extremely large standard

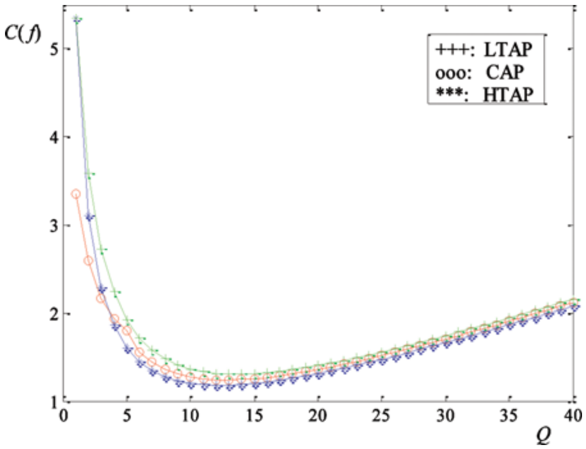

FIGURE 1 $C(f)$ for quantity policies (color figure available online).

deviation of weight per order can cause the accumulated weight, in a short period of time, to go from a small amount to a large amount that exceeds Q. Thus, a shipment is likely to be dispatched right after a large order arrives. Consequently, the mean holding cost can be smaller for the HTAP case. At the optimal threshold Q^*s, the performance measures are given in Table 1.

As shown in Table 1, the mean excess for the HTAP case is significantly larger than that of the other two cases. For LTAP, since demands arrive in a bursty manner, the probability of excess is also significant. On the other hand, there is no excess at all for the CAP case. For all three cases, the optimal quantity is significantly smaller than the excess threshold $Q_o = 20$ (65%, 60%, and 70% for the CAP, HTAP, and LTAP cases, respectively). Yet there is still considerable excess for HTAP and LTAP. Thus, if there is a cost associated with the excess, the mean total cost for HTAP and LTAP may increase significantly.

Example 5.2. Consider the same model as in Example 5.1, except that the dispatch policy is now a hybrid policy. We assume $f(1) = \cdots = f(j_q - 1) = Q = 30$ and $f(j) = 0$ for $j \geq j_q$. Note that $Q = 30$ is a little bit

TABLE 1 Costs associated with the optimal quantity policies

	Q^*	$E[L_c]$	$E[L_w]$	$E[W]$	P_o	$E[O_w]$	$C(f)$
CAP	13	14.05	7.59	5.30	0	0	1.2415
HTAP	12	14.92	8.06	5.07	0.073	1.9815	1.1773
LTAP	14	14.30	7.42	6.09	0.006	0.0117	1.3081

TABLE 2 Costs associated with the optimal hybrid policies.

	j_q^*	$E[L_c]$	$E[L_w]$	$E[W]$	P_o	$E[O_w]$	$C(f)$
CAP	14	13.995	6.498	6.722	0.148	0.537	1.3867
HTAP	16	15.653	7.420	6.416	0.165	2.713	1.2805
LTAP	14	13.993	6.497	6.776	0.125	0.461	1.3922

more than double the optimal quantities obtained in Example 5.1. Again, we set $Q_o = 20$ for the excess level. The average total costs per unit time are computed for $j_q = 1, 2, \ldots, 30$, and the results are plotted in Figure 2.

The optimal j_q is 14, 16, and 14 for the respective three cases. The HTAP case has the smallest mean cost: Its weight per order can be very large, hence the accumulated weight can exceed Q and all that weight be shipped out before the scheduled time j_q. Thus, the HTAP case avoids carrying large inventory, as compared to the other two cases. Details on the individual optimal results are given in Table 2.

The chance of getting an excess, and the mean excess of the HTAP case, are significantly larger than those of the other two cases. If there is a cost associated with an excess weight, HTAP may have a higher cost.

There are cases for which the optimal hybrid policy is better than the optimal quantity policy. As shown by the next example, there are also cases for which a generally-structured policy performs better than both the best quantity policy and the best hybrid policy.

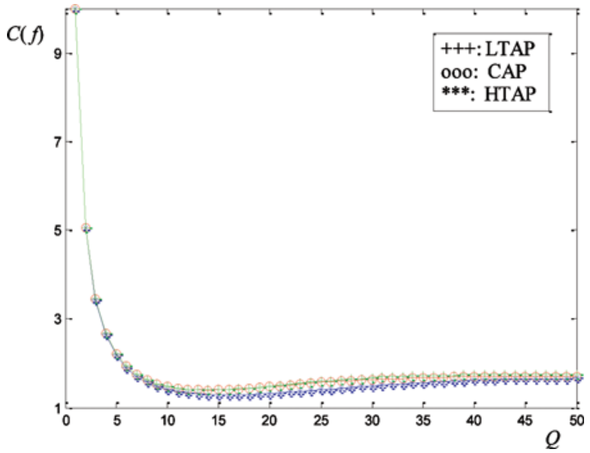

FIGURE 2 $C(f)$ for hybrid policies (color figure available online).

Example 5.3. Consider a model with $K_d = 10$, $h = 0.1$, and weight accumulation process:

$$D_0 = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.8 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 0.2 & 0.1 \\ 0 & 0.1 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0.2 & 0 \\ 0 & 0 \end{pmatrix},$$
$$D_3 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0 \end{pmatrix}, \quad D_4 = \begin{pmatrix} 0.1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{5.3}$$

We set $Q_o = 20$. The optimal quantity policy is $Q* = 10$ with $C(f) = 0.9238$, $E[L_c] = 20.039$, $E[L_w] = 14.543$, $P_o = 0$, and $E[O_w] = 0$. The optimal hybrid policy with $Q = 30$ is $j_q^* = 20$ with $C(f) = 1.003$, $E[L_c] = 19.951$, $E[L_w] = 9.481$, $P_o = 0.115$, and $E[O_w] = 0.589$.

By a manual search, the following generally-structured dispatch policy is found to be better than the respective optimal choice for both the quantity and hybrid policies. For $f(j) = 16$, $1 \le j \le 5$; $f(j) = 15, 6 \le j \le 10$; $f(j) = 10, 11 \le j \le 13$; $f(j) = 9, 14 \le j \le 16$; and $f(j) = 8, 17 \le j \le 20$, we have $C(f) = 0.897$, $E[L_c] = 19.499$, $E[L_w] = 12.781$, $P_o = 0$, and $E[O_w] = 0$. As shown, the generally-structured policy has a smaller mean total cost. This example indicates that the optimal policy may not be a quantity policy nor a time policy. The time elapsed since the last dispatch must be taken into consideration when it comes to cost reduction.

6. CONCLUSIONS

In this article, shipment consolidation has been studied for the case of private carriage, and an arrival process more general than those in the existing literature. The discrete weight of an individual order may be correlated with its arrival time. Efficient algorithms were developed for computing the performance measures for several policies (quantity, time, hybrid policies) commonly used in industrial practice. We did give one example of a more general dispatch policy where the average total cost was better than that of the preceding policies. But it would be interesting to develop algorithms for finding the best *general* dispatch policy to minimize the average total costs.

Moreover, it would be worthwhile to develop analyses and algorithms corresponding to those of the present article, now for the case of common carrier transportation. We have already begun that research. For example, the matrix-analytic methods, and the special QBD structure of the Markov chain introduced for the system, enable us to obtain a distribution for the weight of an individual load. That is crucial for calculating the expected cost of common carriage.

ACKNOWLEDGMENT

The authors thank the referees for their useful comments and suggestions.

REFERENCES

1. Bookbinder, J.H.; Higginson, J.K. Probabilistic modeling of freight consolidation by private carriage. Transportation Research, Part E. **2002**, *38*, 305–318.
2. Çetinkaya, S.; Bookbinder, J.H. Stochastic models for the dispatch of consolidated shipments. Transportation Research, Part B. **2003**, *37*, 747–768.
3. Çetinkaya, S. Coordination of inventory and shipment consolidation decisions: A review of premises, models, and justification. In *Applications of Supply Chain Management and E-Commerce Research*; Springer: New York, 2005.
4. He, Q.M.; Neuts, M.F. Markov chains with marked transitions. Stochastic Processes and Their Applications **1998**, *74*, 37–52.
5. Higginson, J.K.; Bookbinder, J.H. Policy recommendations for a shipment consolidation program. Journal of Business Logistics **1994**, *15*, 87–112.
6. Higginson, J.K.; Bookbinder, J.H. Markovian Decision processes in shipment consolidation. Transportation Science **1995**, *29*, 242–255.
7. Latouche, G.; Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modelling*; ASA & SIAM: Philadelphia, USA, 1999.
8. Latouche, G.; Remiche, M.A.; Taylor, P. Transit Markov arrival processes. The Annals of Applied Probability **2003**, *13*, 628–640.
9. Lucantoni, D.M. New results on the single server queue with a batch Markovian arrival process. Stochastic Models **1991**, *17*, 1–46.
10. Mütlü, F.; Çetinkaya, S.; Bookbinder, J.H. An analytical model for computing the optimal time-and-quantity-based policy for consolidated shipments. IIE Transactions **2010**, *42*, 367–377.
11. Neuts, M.F. A versatile Markovian point process. Journal of Applied Probability **1979**, *16*, 764–79.
12. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models—An Algorithmic Approach*; The Johns Hopkins University Press: Baltimore, 1981.
13. Ülkü, M.A.; Bookbinder, J.H. Policy analysis in shipment consolidation. In *Proceedings of the 26th Turkish National OR/IE Conference*; Kocaeli, Turkey, 2006, pp. 9–12.