Vendor-managed inventory (VMI) is a supply-chain initiative where the supplier is authorized to manage inventories of agreed-upon stock-keeping units at retail locations. The benefits of VMI are well recognized by successful retail businesses such as Wal-Mart. In VMI, distortion of demand information (known as bullwhip effect) transferred from the downstream supply-chain member (e.g., retailer) to the upstream member (e.g., supplier) is minimized, stockout situations are less frequent, and inventory-carrying costs are reduced. Furthermore, a VMI supplier has the liberty of controlling the downstream resupply decisions rather than filling orders as they are placed. Thus, the approach offers a framework for synchronizing inventory and transportation decisions.

In this paper, we present an analytical model for coordinating inventory and transportation decisions in VMI systems. Although the coordination of inventory and transportation has been addressed in the literature, our particular problem has not been explored previously. Specifically, we consider a vendor realizing a sequence of random demands from a group of retailers located in a given geographical region. Ideally, these demands should be shipped immediately. However, the vendor has the autonomy of holding small orders until an agreeable dispatch time with the expectation that an economical consolidated dispatch quantity accumulates. As a result, the actual inventory requirements at the vendor are partly dictated by the parameters of the shipment-release policy in use. We compute the optimum replenishment quantity and dispatch frequency simultaneously. We develop a renewal-theoretic model for the case of Poisson demands, and present analytical results.

1. Problem Context
Contemporary research in supply-chain management relies on an increased recognition that an integrated plan, for the chain as a whole, requires coordinating different functional specialties within a firm (e.g., marketing, procurement, manufacturing, distribution, etc.). Consequently, emphasis on supply-chain coordination has increased in recent years (e.g., see Arntzen et al. 1995, Lee and Billington 1992, Lee et al. 1997, and Tayur et al. 1999). In keeping with this trend, we focus on the coordination efforts aimed at the integration of inventory and transportation.

Pioneered by Wal-Mart, implemented by Fruit of the Loom, Shell Chemical, and others, Vendor-managed inventory (VMI) is an important coordination initiative. In VMI, the vendor assumes responsibility for managing inventories at retailers using advanced online messaging and data-retrieval sys-
tems (Aviv and Federgruen 1998, Parker 1996, Schenck and McInerny 1998). Reviewing the retailer’s inventory levels, the supplier makes decisions regarding the quantity and timing of resupply. This implies that inventory information at the retailer is accessible to the supplier. As a result, the approach is gaining more attention as electronic data interchange (EDI) technology improves and the cost of information sharing decreases.

Although current interest in supply-chain management overlooks certain transportation/distribution issues, substantial savings are realizable by carefully incorporating a shipment strategy with the stock replenishment decisions for VMI systems. This impact is particularly tangible when the shipment strategy calls for a consolidation program where several smaller deliveries are dispatched as a single load, realizing scale economies inherent in transportation. Formally, shipment consolidation refers to the active intervention by management to combine many small shipments/orders so that a larger, hence more economical, load can be dispatched on the same vehicle (Brennan 1981, Hall 1987, Higginson and Bookbinder 1995). The main motivation behind a consolidation program is to take advantage of the decreased per unit freight costs due to economies of scale associated with transportation.

Efficient use of transportation resources is particularly important for successful implementation of just-in-time (JIT) procurement systems and VMI systems. JIT is widely recognized as a way of minimizing inventories. However in JIT, inventory is often reduced at the expense of frequent shipments of smaller loads (Arcelus and Rowcroft 1991, Gupta and Bagchi 1987). Applied carefully, a consolidation program avoids the supply-chain members trading off inventory costs for substantial transportation costs in a JIT environment (Popken 1994). Concern over the interaction between transportation and inventory costs has long been discussed in the JIT literature (Yano and Gerchak 1989). However, in the context of VMI, shipment-consolidation is an unexplored area.

For VMI applications, the supplier is empowered to control the timing and quantity of downstream resupply decisions. Thus, the supplier has more freedom to consolidate resupply shipments over time and geographical regions. Full vehicles are more likely to be dispatched, transportation scale economies are easier to achieve, and there is ample opportunity to synchronize the inventory and transportation decisions. Our research builds on these practical motivations concerning VMI systems, and our objective is to provide analytical models for improving inventory and transportation decisions.

Here, we develop an integrated stock replenishment and delivery scheduling model for a VMI supplier. We consider the case where the vendor uses a certain kind of \((s,S)\) policy for replenishing its inventory, and a time-based, shipment-consolidation policy for delivering customer demands. Our additional assumptions regarding the inventory system under consideration follow:

- The vendor observes a sequence of random demands from a group of retailers located in a given geographical region as they are realized. Ideally, these demands should be shipped immediately. However, we consider the case where the vendor has the liberty of not delivering small orders until an agreeable dispatch time, with the expectation that an economical dispatch quantity accumulates. This shipment release policy is classified as a time-based policy, and it is explained in detail in §2.
- Retailers are willing to wait at the expense of waiting costs, and the vendor assures customer satisfaction by imposing a latest dispatch time. The main motivation of the vendor for such a delivery policy is to dispatch larger loads benefiting the economies of scale inherent in transportation. On the other hand, the retailers agree to wait when their shelf space is limited for carrying extra stock, or else carrying inventory for certain items is not desirable. This situation is particularly common for retail stores making catalog sales.
- We say that a new “shipment-consolidation cycle” begins each time a dispatch decision is taken. Thus, all orders arriving during a consolidation cycle are combined to form a large dispatch quantity. Once a dispatch decision is taken, the entire load is shipped, i.e., all outstanding demands are delivered. If, at the time of dispatch, the on-hand inventory is insufficient to clear all outstanding demands, then the vendor

---

Management Science / Vol. 46, No. 2, February 2000
immediately replenishes its stock from an external source with ample supply.

Under the above assumptions, our problem is to compute an optimum replenishment quantity and dispatch frequency for the case of random demands. Our approach is to minimize total procurement, transportation, inventory carrying, and waiting costs simultaneously while satisfying customer requirements on a timely basis. We provide a renewal theoretic model based on this approach and obtain analytical results.

A review of the relevant literature is provided in the next section. We note that the literature on economic inventory/transportation decisions is abundant. In the interest of brevity, our discussion reviews only those models most closely related to our research in VMI systems. The characteristics of the problem under consideration are discussed in §3. The problem formulation is developed in §4, a detailed analysis for computing the optimal solution is presented in §5, and some numerical examples are furnished in §6. In §7, we provide a point of comparison for the integrated policy presented in the paper and an immediate delivery policy. Finally, §8 includes our concluding comments.

2. Relevant Literature

In order to specify a shipment release schedule, most firms implement an operating routine whereby a selected dispatching policy is employed each time a demand arrives (Abdelwahab and Sargious 1990). Typically, relevant criteria for selecting an operating routine involve customer satisfaction as well as cost minimization (Newbourne and Barrett 1972). Some operational issues in managing shipment release systems are similar to those encountered in inventory control. Two fundamental questions of shipment release scheduling are i) when to dispatch a vehicle so that service requirements are met, and ii) how large the dispatch quantity should be so that scale economies are realized. It is worth noting that these two questions involve shipment-consolidation across time since a consolidated load accumulates by holding shipments over a period. This practice is also known as temporal consolidation.

Identification of practical operating routines for temporal consolidation has received significant attention (e.g., Burns et al. 1985, Hall 1987, Russell and Krajewski 1991). Some researchers use simulation as a modeling tool (e.g., Closs and Cook 1987, Higginson and Bookbinder 1994, Masters 1980) whereas others (e.g. Higginson and Bookbinder 1995, Gupta and Bagchi 1987, Powell 1985, Stidham 1977, Brennan 1981, Minkoff 1993) provide models using analytical approaches such as Markov decision processes, queueing theory, and dynamic programming. However, existing literature on temporal consolidation ignores joint stock replenishment and delivery-scheduling decisions that arise in the context of VMI systems.

In general terms, existing literature identifies two different types of temporal consolidation routines (Higginson and Bookbinder 1994). These are i) time-based dispatch policies, and ii) quantity-based dispatch policies. A time-based policy ships an accumulated load (clears all outstanding demands) every T periods. A quantity-based policy ships an accumulated load when an economic freight quantity is available. Under a time-based policy, the actual dispatch quantity is a random variable for the case of stochastic demands. Therefore, transportation scale economies may not be realized for some demand instances. However, a time-based policy assures that each demand is dispatched by a predetermined shipping date. Hence, when a demand is placed the shipper can quote a specific delivery time with certainty. In contrast, under a quantity-based policy, dispatch time is a random variable whereas dispatch quantity assures scale economies. Apparently, a time-based policy is more desirable for satisfying timely customer service requirements at the expense of sacrificing some scale economies.

In recent years, time-based shipment consolidation policies have become a part of transportation contracts between partnering supply-chain members. These contracts are also known as time definite delivery (TDD) agreements that are common between third-party logistics service providers and their partnering manufacturing companies. In a representative practical situation, a third-party logistics company provides
warehousing and transportation for a manufacturer and guarantees TDD for outbound deliveries to the customers. Such an arrangement is particularly useful for effective VMI. Similarly, the third party may provide TDD for inbound deliveries for the manufacturer itself, and such an arrangement is useful for effective JIT manufacturing. TDD contracts for inbound and outbound deliveries are particularly common in the Texas personal computer industry.

The problem of interest in this study concerns the case where the vendor adopts a time-based policy for shipment release timing. Along with the optimal replenishment quantity for the vendor, the shipment frequency, $T$, is also a decision variable. The objective is to optimize the inventory and dispatch decisions simultaneously. We develop a renewal theoretic optimization model and obtain a closed form solution for the particular case of Poisson demands. We remark that current interest in VMI focuses on measuring the value of information through an analysis of the operational characteristics of VMI systems (e.g., Aviv and Federgruen 1998, Bourland et al. 1996, Cheung and Lee 1998, Lee et al. 1997), whereas our model is aimed at optimizing inventory and transportation decisions.

It is worth noting that inventory lot-sizing models where transportation costs are considered explicitly are also related to our study. To the best of our knowledge, the efforts in this area are mainly directed towards deterministic modeling. In particular, deterministic joint quantity and freight discount problems have received significant attention. As an example, Tersine and Barman (1991) model such a dual discount situation for the purpose of lot-size optimization. Other deterministic lot-sizing and transportation papers include Gupta (1992), Hahn and Yano (1995a, 1995b), Lee (1986, 1989), and Van Eijs (1994). In addition, Popken (1994) studies a deterministic multi-attribute, multicommodity flow problem with freight consolidation, Tyworth (1992) develops a framework for analyzing inventory and transportation trade-offs, and Henig et al. (1997) explore the joint optimization of inventory and supply contract parameters in a stochastic setting. The literature on the inventory-routing problem is also related to our research (e.g., see Kleywegt et al. 1998). However, these earlier papers do not consider the effects of stochastic temporal consolidation that arise in the context of VMI. Thus, from a practical point of view, there is still a need for analytical models that take into account freight-consolidation in a stochastic setting.

3. Problem Characteristics
To set the stage for a mathematical formulation, we present an example illustrated in Figure 1, in which $M$ is a manufacturer and $V$ is a vendor/distributor. A group of retailers ($R_1, R_2, \text{etc.}$) located in a given geographical region has random demands with identical sizes, and these can be consolidated in a larger load before a delivery is made to the region. That is, demands are not satisfied immediately, but rather are shipped in batches of consolidated loads. As a result, the actual inventory requirements at $V$ are specified by the dispatching policy in use, and consolidation and inventory decisions at $V$ should not be made in isolation from each other. In this example, the total cost for the vendor includes procurement and inventory carrying costs at $V$, the cost of waiting associated with ordered but not-yet-delivered demand items, and transportation costs for shipments from $V$ to the region.

We take into account for the following cost parameters:

- $A_R$: Fixed cost of replenishing inventory
- $c_R$: Unit procurement cost
- $h$: Inventory carrying cost per unit per unit time
- $A_D$: Fixed cost of dispatching

**Figure 1** Consolidation in VMI
Here, costs $A_d$, $c_d$, and $h$ are the traditional model parameters associated with inventory replenishment problems. Parameters $A_d$ and $c_d$ correspond to the delivery costs using a private fleet. In the classical inventory models, $A_d$ and $c_d$ are sunk costs and need not be modeled. This is because the traditional literature assumes that demands are satisfied as they arrive. However, if shipments are consolidated, then delivery costs play an important role for balancing the trade-off between scale economies associated with transportation and customer waiting. Therefore, we should take into account for $A_d$ and $c_d$ explicitly. For the problem of interest, $w$ represents a loss-of-goodwill penalty as well as an opportunity loss in postponed receipt of revenues. If demands are delivered as they arrive (without consolidation), then no waiting costs accumulate.

Assuming that dispatch decisions (at $V$) are made on a recurrent basis, we utilize renewal theory to obtain an optimal solution for the problem of interest. We suppose that demands (from customers located in a given region) form a stochastic process with interarrival times $X_n: n = 1, 2, \ldots$. We consider the case where $X_n \geq 0$, $n = 1, 2, \ldots$ are independent and identically distributed (i.i.d.) according to $F(\cdot)$ where $F(0) < 1$. Letting $S_0 = 0$ and $S_n = \sum_{j=1}^{n} X_j$, we define

$$N(t) = \sup\{ n: S_n \leq t \}.$$  

It follows that $N(t)$ is a renewal process that registers the number of demand orders placed by time $t$. A realization of $N(t)$ is depicted in Figure 2.

Adopting a time-based consolidation policy, a dispatch decision is taken every $T$ time units (e.g., days). In turn, maximum holding time $T$ correspondingly represents the length of a shipment-consolidation cycle. A realization of the consolidation process, denoted by $L(t)$, is depicted in Figure 3. Observe that $L(t)$ represents the size of the accumulated load, i.e., number of outstanding demands, at time epoch $t$.

Let $I(t)$ denote the inventory level at time $t$ and $Q$ denote the inventory level immediately after a replenishment order arrives. We assume that inventory replenishment lead time is negligible, and a replenishment order is placed only if the outstanding orders cannot be cleared using the on-hand inventory. In this case, it is sufficient to review inventory immediately before a dispatch decision, i.e., at time points $T, 2T, \ldots$. Then,

1. $I(t)$ and $L(t)$, $t = T, 2T, \ldots$, are observed.
2. The vendor employs a special kind of $(s, S)$ policy with $s = 0$ and $S = Q$. For our problem, an $(s, S)$ policy with $s = 0$ is appropriate since the
replenishment lead time at the vendor is negligible. Thus, there is no need to place an order if \( I(t) \geq 0 \) immediately after a shipment is dispatched. Assuming that the vendor replenishes its stock from a manufacturer with ample supply, \( S = Q \) is the order up-to-level after meeting all the demand. Therefore, \( Q \) is called the target inventory level. If \( I(t) < L(t) \), then a replenishment order quantity of \( Z(t) \) is placed where

\[
Z(t) = \begin{cases} 
Q + L(t) - I(t), & \text{if } I(t) < L(t), \\
0, & \text{if } I(t) \geq L(t).
\end{cases}
\]  

(1)

Here, the inventory problem is to compute the optimal value of \( Q \).

3. Upon the receipt of \( Z(t) \), a load containing \( L(t) \) units is dispatched instantaneously.

4. A new shipment-consolidation cycle begins with \( Y(t) \) units of inventory where

\[
Y(t) = \begin{cases} 
Q, & \text{if } I(t) < L(t), \\
I(t) - L(t), & \text{if } I(t) \geq L(t).
\end{cases}
\]

That is, \( Y(t), t = T, 2T, \ldots \) denotes the order-up-to-level for replenishing inventory.

A realization of the inventory process \( I(t) \) is illustrated in Figure 4, where we assume \( Q = 4 \). For the realization of the demand process given in Figure 2, the size of the consolidated load in the first consolidation cycle is 3, i.e., \( L(T) = 3 \). Knowing that \( I(T) = 4 > L(T) \), a load containing 3 units is dispatched at time \( T \) where \( Z(T) = 0 \). The second consolidation cycle begins with 1 unit of inventory, i.e., \( Y(T) = 1 \). As illustrated in Figure 3, a load of 4 units accumulates during the second consolidation cycle, i.e., \( L(2T) = 4 \). However, \( L(2T) = 1 < L(2T) \) so that an order of size \( Z(2T) = Q + L(2T) - I(2T) = 4 + 4 - 1 = 7 \) units is placed and received instantaneously. Immediately after the order of \( Z(2T) = 7 \) units is received, a dispatch decision for delivering the accumulated load of \( L(2T) = 4 \) units is taken. Thus, the third consolidation cycle begins with the target inventory level of \( Q = 4 \) units.

Our problem is to compute the optimal \( Q \) and \( T \) values simultaneously. Therefore, our model is also related to inventory problems with periodic audits where computing the length of a review period is of interest (see Flynn and Gartska 1990). However, unlike in the traditional inventory literature, we take into account outbound transportation costs and the effects of shipment consolidation explicitly.

Let us recall that \( L(t) \) represents the size of the accumulated load, i.e., number of outstanding demands, at time epoch \( t \). The consolidation system is cleared and a new shipment-consolidation cycle begins every \( T \) time units. In turn, \( L(jT), j = 1, 2, \ldots, \) is a sequence of random variables representing the dispatch quantities. Keeping this observation in mind, we define

\[
N_j(T) = L(jT), \quad j = 1, 2, \ldots.
\]

Observe that the sequence \( N_j(T), j = 1, 2, \ldots \) symbolizes the demand process realized by the inventory system under the time-based dispatching policy in use, whereas \( X_n, n = 1, 2, \ldots \) is the actual demand process (see Figures 2, 3, and 4). The process \( N_j(T), j = 1, 2, \ldots \) is a function of \( T \), and thus the actual inventory requirements at the vendor are established by the parameter of the shipment-consolidation policy in use.

If \( N(t) \) is a Poisson process, then the random variables \( N_j(T), j = 1, 2, \ldots, \) are i.i.d., each having the
same distribution as the random variable \( N(T) \). It is worth noting that for other renewal processes \( N_j(T), j = 1, 2, \ldots \), are not necessarily i.i.d., and this is a major source of difficulty for the problem of interest. Obtaining analytical results for general renewal processes seems to be rather challenging if not impossible. Here, we focus on the case of Poisson demand for analytical tractability.

4. Problem Formulation

A replenishment cycle is defined as the time interval between two consecutive replenishment decisions. Under the assumption of Poisson demands, \( I(t) \) can be split into i.i.d. replenishment cycles, and thus \( I(t) \) is a regenerative process. The regeneration points are the epochs at which the target inventory level \( Q \) is reached (i.e., the time epochs immediately after the replenishment decisions). Consequently, we may use the renewal reward theorem for computing the expected costs associated with the inventory system under consideration. Let \( C(Q, T) \) denote the expected long-run average cost. The renewal reward theorem implies that

\[
C(Q, T) = \frac{E[\text{Replenishment Cycle Cost}]}{E[\text{Replenishment Cycle Length}]}.
\]

(2)

Once an expression of \( C(Q, T) \) is obtained, the problem reduces to

\[
\min \ C(Q, T) \quad \text{s.to} \quad Q \geq 0, \quad T \geq 0.
\]

As we have already mentioned, the consolidation system is cleared and a new consolidation cycle is started every \( T \) time units. In turn, a replenishment cycle includes at least one shipment-consolidation cycle. We recall that inventory is replenished when the cumulative demand realized by the inventory system exceeds \( Q \), and we define

\[
K = \inf \left\{ k : \sum_{j=1}^{k} N_j(T) > Q \right\}.
\]

(3)

By definition, \( K \) is a random variable representing the number of dispatch decisions within a given inventory replenishment cycle. It follows that the length of an inventory replenishment cycle is \( KT \), and thus

\[
E[\text{Replenishment Cycle Length}] = E[K]T. \quad (4)
\]

Since \( K \) is a positive random variable, its expected value is given by (Taylor and Karlin 1998, p. 44)

\[
E[K] = \sum_{k=1}^{\infty} P\{K \geq k\}.
\]

Expression (3) suggests that

\[
\{K \geq k\} \sim N\left(\sum_{j=1}^{k-1} N_j(T) \leq Q\right). \quad (5)
\]

Let \( G(\cdot) \) denote the distribution function of \( N(T) \) and \( G^{(k)}(\cdot) \) denote the \( k \)-fold convolution of \( G(\cdot) \). Then the above relation leads to

\[
P\{K \geq k\} = G^{(k-1)}(Q)
\]

so that

\[
E[K] = \sum_{k=1}^{\infty} G^{(k-1)}(Q). \quad (6)
\]

The assumed cost structure implies that, in our formulation, \( E[\text{Replenishment Cycle Cost}] \) consists of 1. inventory replenishment costs; 2. delivery costs; 3. inventory carrying costs; and 4. customer waiting costs. Next, we discuss the computation of each of these cost items.

4.1. Expected Inventory Replenishment Costs per Replenishment Cycle

Under the assumed replenishment cost structure,

\[
E[\text{Inventory replenishment costs per cycle}] = A_R + c_R E[\text{Order quantity}].
\]

Given the ordering policy in (1), expected order quantity is equal to the expected total demand within a replenishment cycle (see Figure 4). Thus,

\[
E[\text{Order quantity}] = E\left[ \sum_{j=1}^{K} N_j(T) \right].
\]
Noting that $K$ is a stopping time for the sequence $N_j(T)$, $j = 1, 2, \ldots$, we have

$$E[\text{Inventory replenishment costs per cycle}] = A_k + c_\nu E[K]E[N(T)].$$  \hfill (7)

### 4.2. Expected Delivery Costs per Replenishment Cycle

Recalling that all outstanding demands are delivered at the time of a dispatch decision, and $K$ is the number of dispatch decisions within a given replenishment cycle,

$$E[\text{Delivery costs per cycle}] = A_pE[K] + c_\nu E[K]E[N(T)].$$  \hfill (8)

### 4.3. Expected Inventory Carrying Costs per Replenishment Cycle

The characteristics of the inventory system (see Figure 4) under consideration imply that

$$I(t) = \begin{cases} Q, & \text{if } 0 \leq t \leq T, \\ Q - N_1(T), & \text{if } T < t \leq 2T, \\ \vdots & \vdots \\ Q - \sum_{j=1}^{K-1} N_j(T), & \text{if } (K-1)T < t \leq KT. \end{cases}$$

Considering an inventory holding cost of $h$/unit/unit-time,

$$E[\text{Inventory carrying costs per cycle}] = hE\left[ \int_0^{KT} I(t)dt \right].$$

Computing the expected inventory carrying costs per replenishment cycle using the above expression is messy. In order to simplify this problem, we define

$$H(Q, T) = E\left[ \int_0^{KT} I(t)dt \right],$$

and prove the following proposition.

**Proposition 1.** Let $g(\cdot)$ denote the probability mass function of $N(T)$, and $g^{(k)}(\cdot)$ denote the $k$-fold convolution of $g(\cdot)$. Then

$$H(Q, T) = TQ + T \sum_{i=0}^{Q} (Q-i)m_g(i)$$  \hfill (9)

where

$$m_g(i) = \sum_{k=1}^{\infty} g^{(k)}(i),$$  \hfill (10)

i.e., $m_g(\cdot)$ is the renewal density associated with $g(\cdot)$.

**Proof.** See the Appendix.

It follows from Proposition 1 that

$$E[\text{Inventory carrying costs per cycle}] = hH(Q, T)$$

$$= hTQ + hT \sum_{i=0}^{Q} (Q-i)m_g(i).$$  \hfill (11)

### 4.4. Expected Customer Waiting Costs per Replenishment Cycle

Figure 5 illustrates the accumulation of a consolidated weight and inventory costs. It is easy to observe that

$$E[\text{Waiting cost per consolidation cycle}] = wE\left[ \sum_{n=2}^{N(T)} (j-1)X_n + N(T)\alpha(T) \right]$$  \hfill (12)

where $\alpha(T) = T - S_{N(T)}$ denotes the age of $N(t)$ at $T$. Alternatively, we can write

$$E[\text{Waiting cost per consolidation cycle}] = wE[(T - S_1) + (T - S_2) + \ldots + (T - S_{N(T)})]$$

$$= wE\left[ N(T)T - \sum_{n=1}^{N(T)} S_n \right].$$  \hfill (13)

Obtaining an explicit expression of expected waiting cost per consolidation cycle directly from Equations (12) or (13) requires some effort. However, for our purposes, the problem can be simplified a great deal by letting

$$W(T) = E\left[ N(T)T - \sum_{n=1}^{N(T)} S_n \right],$$
and using the following proposition.

**Proposition 2.**

\[ W(T) = v(T) + \int_0^T v(T - t)dM_f(t) \]  

where

\[ v(T) = \int_0^T (T - t)dF(t), \]

\[ M_f(t) = \sum_{n=1}^\infty F^{(n)}(t) = E[N(t)]. \]

**Proof.** See the Appendix.

In conclusion,

\[
E[\text{Waiting costs per replenishment cycle}]
= wE[K]E[\text{Waiting costs per consolidation cycle}]
= wE[K]W(T)
= wE[K]v(T) + wE[K] \int_0^T v(T - t)dM_f(t). \]  

\[ (14) \]

\[ (15) \]

\[ (16) \]

Given the type of demand process, we aim to obtain an explicit expression of \( C(Q, T) \) and determine the optimal \((Q, T)\) pair. To this end, we consider additional properties of Poisson processes.

### 5.1. An Explicit Expression of the Long-Run Average Cost

If the demand arrivals, \( N(t) \), follow a Poisson process with parameter \( \lambda t \) then

\[ N(T) \text{ is a Poisson random variable with parameter } \lambda T, \]

\[ G \text{ is a Poisson distribution with parameter } \lambda T. \]

In turn,

\[ E[N(T)] = \lambda T. \]  

\[ (17) \]

**Proof.** See the Appendix.

In conclusion,

\[ E[\text{Waiting costs per replenishment cycle}]
= wE[K]E[\text{Waiting costs per consolidation cycle}]
= wE[K]W(T)
= wE[K]v(T) + wE[K] \int_0^T v(T - t)dM_f(t). \]  

\[ (14) \]

\[ (15) \]

\[ (16) \]

Given the type of demand process, we aim to obtain an explicit expression of \( C(Q, T) \) and determine the optimal \((Q, T)\) pair. To this end, we consider additional properties of Poisson processes.

### 5.1. An Explicit Expression of the Long-Run Average Cost

If the demand arrivals, \( N(t) \), follow a Poisson process with parameter \( \lambda t \) then

\[ N(T) \text{ is a Poisson random variable with parameter } \lambda T, \]

\[ G \text{ is a Poisson distribution with parameter } \lambda T. \]

In turn,

\[ E[N(T)] = \lambda T. \]  

\[ (17) \]

**Proof.** See the Appendix.

In conclusion,

\[ E[\text{Waiting costs per replenishment cycle}]
= wE[K]E[\text{Waiting costs per consolidation cycle}]
= wE[K]W(T)
= wE[K]v(T) + wE[K] \int_0^T v(T - t)dM_f(t). \]  

\[ (14) \]

\[ (15) \]

\[ (16) \]

Given the type of demand process, we aim to obtain an explicit expression of \( C(Q, T) \) and determine the optimal \((Q, T)\) pair. To this end, we consider additional properties of Poisson processes.

### 5.1. An Explicit Expression of the Long-Run Average Cost

If the demand arrivals, \( N(t) \), follow a Poisson process with parameter \( \lambda t \) then

\[ N(T) \text{ is a Poisson random variable with parameter } \lambda T, \]

\[ G \text{ is a Poisson distribution with parameter } \lambda T. \]

In turn,

\[ E[N(T)] = \lambda T. \]  

\[ (17) \]

**Proof.** See the Appendix.

In conclusion,

\[ E[\text{Waiting costs per replenishment cycle}]
= wE[K]E[\text{Waiting costs per consolidation cycle}]
= wE[K]W(T)
= wE[K]v(T) + wE[K] \int_0^T v(T - t)dM_f(t). \]  

\[ (14) \]

\[ (15) \]

\[ (16) \]

Given the type of demand process, we aim to obtain an explicit expression of \( C(Q, T) \) and determine the optimal \((Q, T)\) pair. To this end, we consider additional properties of Poisson processes.

### 5.1. An Explicit Expression of the Long-Run Average Cost

If the demand arrivals, \( N(t) \), follow a Poisson process with parameter \( \lambda t \) then

\[ N(T) \text{ is a Poisson random variable with parameter } \lambda T, \]

\[ G \text{ is a Poisson distribution with parameter } \lambda T. \]

In turn,

\[ E[N(T)] = \lambda T. \]  

\[ (17) \]

**Proof.** See the Appendix.

In conclusion,

\[ E[\text{Waiting costs per replenishment cycle}]
= wE[K]E[\text{Waiting costs per consolidation cycle}]
= wE[K]W(T)
= wE[K]v(T) + wE[K] \int_0^T v(T - t)dM_f(t). \]  

\[ (14) \]

\[ (15) \]

\[ (16) \]

Given the type of demand process, we aim to obtain an explicit expression of \( C(Q, T) \) and determine the optimal \((Q, T)\) pair. To this end, we consider additional properties of Poisson processes.

### 5.1. An Explicit Expression of the Long-Run Average Cost

If the demand arrivals, \( N(t) \), follow a Poisson process with parameter \( \lambda t \) then

\[ N(T) \text{ is a Poisson random variable with parameter } \lambda T, \]

\[ G \text{ is a Poisson distribution with parameter } \lambda T. \]

In turn,

\[ E[N(T)] = \lambda T. \]  

\[ (17) \]

**Proof.** See the Appendix.

In conclusion,
results in

Also, utilizing Equations (18) and (19) in Equation (14) easily to show that

\[ m_s(i) = \sum_{k=1}^{\infty} \frac{(k\lambda T)^i e^{-k\lambda T}}{i!}. \]  \hspace{1cm} (22)

Utilizing Equation (21) in Equation (6) yields

\[ E[K] = \sum_{k=1}^{\infty} \sum_{i=0}^{Q} \frac{(k\lambda T)^i e^{-k\lambda T}}{i!}. \]  \hspace{1cm} (23)

The above information is sufficient to compute \( E[K] \), \( H(Q, T) \), and \( W(T) \), so that \( C(Q, T) \) in Equation (16) can be evaluated. Although useful, a simple closed form expression for \( E[K] \) is not obtainable since the right hand side of Equation (23) does not simplify in a straightforward fashion. Similarly, without a simplified closed form formula for \( m_s(i) \), function \( H(Q, T) \), and hence function \( C(Q, T) \), need to be evaluated numerically. Hence the optimal solution can only be computed numerically. Fortunately, these drawbacks for further analytical results can be overcome by using the subsequent approximations for \( E[K] \) and \( m_s(i) \).

**Proposition 3.** A continuous approximation for \( K \) is provided by an Erlang random variable with scale parameter \( \lambda T \) and shape parameter \( Q \). Thus,

\[ E[K] = \frac{Q + 1}{\lambda T}. \]  \hspace{1cm} (24)

**Proof.** See the Appendix.

**Proposition 4.** An approximation for \( m_s(i) \) is given by

\[ m_s(i) = \frac{1}{\lambda T}. \]  \hspace{1cm} (25)

**Proof.** See the Appendix.

Substituting Equation (25) into Equation (9), it is easy to show that

\[ H(Q, T) = TQ + \frac{Q(Q + 1)}{2\lambda}. \]  \hspace{1cm} (26)

Also, utilizing Equations (18) and (19) in Equation (14) results in

\[ W(T) = \frac{\lambda T^2}{2}. \]  \hspace{1cm} (27)

At last, if we compute Equation (16) using Equations (17), (24), (26), and (27), then we obtain

\[ C(Q, T) = \frac{A_R\lambda}{Q} + c_R\lambda + \frac{h\lambda T Q}{Q + 1} + \frac{hQ}{2} + \frac{A_D}{T} + c_D\lambda + \frac{w\lambda T}{2}. \]  \hspace{1cm} (28)

For the sake of simplicity, we substitute \( Q + 1 = \bar{Q} \) in the above. Then, the expected long-run average cost is expressed as

\[ C(\bar{Q}, T) = \frac{A_R\lambda}{\bar{Q}} + c_R\lambda + \frac{h\lambda T(\bar{Q} - 1)}{\bar{Q}} + \frac{h(\bar{Q} - 1)}{2} + \frac{A_D}{T} + c_D\lambda + \frac{w\lambda T}{2}. \]  \hspace{1cm} (29)

Thus, our problem is given by

\[ \min \quad C(\bar{Q}, T) \quad \text{s.to} \quad \bar{Q} \geq 1 \quad T \geq 0. \]

**5.2. The Solution**

If we treat \( \bar{Q} \) as a continuous variable, then the Hessian of \( C(\bar{Q}, T) \), denoted by \( C \), is computed as

\[ [C] = \begin{bmatrix} [2\lambda(A_R - hT)] / \bar{Q}^3 & h\lambda / \bar{Q}^2 & 2A_D / \bar{Q}^3 \\ h\lambda / \bar{Q}^2 & h^2\lambda^2 / \bar{Q}^2 & 2A_D / \bar{Q}^3 \\ 2A_D / \bar{Q}^3 & 2A_D / \bar{Q}^3 & 2A_D / \bar{Q}^3 \end{bmatrix}. \]

Consequently, the determinant of \( [C] \) is given by

\[ \det[C] = 4\lambda A_D(AR - hT)\bar{Q} - h^2\lambda^2\bar{Q}T^3 / \bar{Q}^4T^3. \]

Using \( \det[C] \), it is straightforward to show that \( C(\bar{Q}, T) \) is convex in \( T \), and not necessarily convex in \( \bar{Q} \). The reason for this complication is due to the term

\[ \frac{h\lambda T(\bar{Q} - 1)}{\bar{Q}} \]

in Equation (28), since all other terms are jointly convex in \( \bar{Q} \) and \( T \).

Let \((\bar{Q}^*, T^*)\) denote the solution of Equation (29). Observe that the optimal solution of our original
problem is given by \((Q^* - 1, T^*)\) since we let \(Q = \bar{Q} - 1.\) The necessary conditions for an optimal solution for Equation (29) yield
\[
\bar{Q} = \sqrt{\frac{2A_R\lambda}{h} - 2\lambda T}, \quad (30)
\]
and thus \(Q^*\) and \(T^*\) are computed by solving Equations (30) and (31) iteratively. However, there is no guarantee that a solution obtained by solving these two equations gives \(Q^*\) and \(T^*\) since the objective function is not jointly convex in the decision variables. Although this observation may seem burdensome, our following analysis simplifies the optimization procedure a great deal.

If we substitute Equation (31) into Equation (28), \(C(\bar{Q}, T)\) reduces to \(C(\bar{Q}).\) After a few algebraic manipulations, it is easy to show that
\[
C(\bar{Q}) = \frac{A_R\lambda}{Q} + \frac{h\bar{Q}}{2} + \sqrt{\frac{2A_D\lambda[2h(\bar{Q} - 1) + w\bar{Q}]}{Q}} + c_R\lambda + c_D\lambda - \frac{h}{2}. \quad (32)
\]

Let us define
\[
C_1(\bar{Q}) = \frac{A_R\lambda}{Q} + \frac{h\bar{Q}}{2}, \quad (33)
\]
\[
C_2(\bar{Q}) = \sqrt{\frac{2A_D\lambda[2h(\bar{Q} - 1) + w\bar{Q}]}{Q}}. \quad (34)
\]

Also let \(C'(\bar{Q}), C'_1(\bar{Q}),\) and \(C'_2(\bar{Q})\) denote the first derivatives of functions \(C(\bar{Q}), C_1(\bar{Q}),\) and \(C_2(\bar{Q}),\) respectively. It follows that
\[
C'(\bar{Q}) = C'_1(\bar{Q}) + C'_2(\bar{Q}), \quad (35)
\]
and \(Q^*\) is one of the solutions of
\[
C'_1(\bar{Q}) + C'_2(\bar{Q}) = 0. \quad (36)
\]

**Theorem 1.** If there exists a solution for (36) over \([1, +\infty)\) then it is unique.

**Proof.** See the Appendix.

Observe that, if \(-C'_1(1) < C'_1(1),\) then these two functions do not intersect over \([1, +\infty).\) In this case, function \(C(\bar{Q})\) does not have a stationary point over the feasible region of our problem. Let us analyze the conditions under which this is true. Using Equations (33) and (34), it is easy to show that
\[
C'_1(1) = -A_R\lambda + \frac{h}{2},
\]
\[
C'_2(1) = -h \sqrt{\frac{2A_D\lambda}{w}}. \quad (37)
\]

Substituting the above two equations into Equation (35), we obtain
\[
C'(1) = -A_R\lambda + \frac{h}{2} - h \sqrt{\frac{2A_D\lambda}{w}}. \quad (38)
\]

Observe that, if
\[
\frac{2A_D\lambda}{w} > \frac{A_R\lambda}{h} - \frac{1}{2}, \quad (39)
\]
then \(-C'_1(1) < C'_1(1),\) i.e., \(-C'_1(\bar{Q})\) and \(C'_1(\bar{Q})\) do not intersect over \([1, +\infty).\) At the same time, if Equation (38) holds, then \(C'(1) > 0,\) i.e., \(C(\bar{Q})\) is increasing at 1. Using Expression (32), we can show that \(C(\bar{Q})\) is increasing as \(\bar{Q}\) goes to infinity. In turn, if Equation (38) holds, then \(C(\bar{Q})\) is an increasing function over \([1, +\infty).\) The following theorem utilizes these observations and states the optimal solution.

**Theorem 2.** Provided that
\[
\frac{2A_D\lambda}{w} \leq \frac{A_R\lambda}{h} - \frac{1}{2}, \quad (39)
\]
\((Q^*, T^*)\) is the unique solution of (30) and (31). Otherwise, \(Q^* = 1\) and \(T^* = \sqrt{2A_D\lambda/w}.
\]

**Proof.** See the Appendix.

Let us recall that the optimal solution of our original problem is specified by \((Q^* - 1, T^*).\) If Equation (39) is not satisfied, then Theorem 2 states that \(Q^* - 1 = 0.\) In this case,
\begin{itemize}
  \item the optimal target inventory level is zero,
  \item all orders arriving during the periods of \(T^* = \sqrt{2A_D\lambda/w}\) time units are consolidated, and thus
\end{itemize}
Observe that the formula for $Q$ are sufficiently close. Indeed, if $h$ is extremely high, then a target inventory level of 0 makes sense.

Knowing that $(Q^*, T^*)$ is obtained by solving Equations (30) and (31) iteratively, it is straightforward to compute the optimal solution. The following lemma brackets the value of $Q^*$ so that the numerical problem is simplified further.

**Lemma 1.** If Equation (39) holds, then

$$\sqrt{\frac{2A_D}{h}} - 2\lambda \sqrt{\frac{2A_D}{h}} \leq Q^* \leq \sqrt{\frac{2A_R}{h}}.$$ 

**Proof.** See the Appendix.

Finally, let us note that for large $\bar{Q}$,

$$\frac{h\lambda T(\bar{Q} - 1)}{\bar{Q}} \geq 3h\lambda T,$$

and thus

$$C(\bar{Q}, T) \geq \frac{A_R\lambda}{\bar{Q}} + c_R\lambda + h\lambda T$$

$$+ \frac{h(\bar{Q} - 1)}{2} + \frac{A_D}{T} + c_D\lambda + \frac{w\lambda T}{2}.$$ (40)

Furthermore, it can be easily proved that the above is jointly convex in $\bar{Q}$ and $T$, and its global minimum is reached at

$$\bar{Q} = \sqrt{\frac{2A_R\lambda}{h}}, \quad T = \sqrt{\frac{2A_D}{\lambda(w + 2h)}}.$$ (41)

Observe that the formula for $\bar{Q}$ in Equation (41) is the standard EOQ formula. For practical reasons, we believe that Equation (41) provides a good approximation for the solution of Equations (30) and (31), and our numerical illustrations support this argument, i.e., the optimal solution and the solution in Equation (41) are sufficiently close.

It is worth noting that if $\bar{Q}$ is computed without considering the cost implications of shipment consolidation, i.e., by minimizing the first four terms of Equation (28), then the optimal $\bar{Q}$ is simply given by the EOQ formula as in Equation (41). In a similar fashion, if $T$ is computed without considering inventory replenishment and carrying costs, i.e., by minimizing the last three terms of Equation (28) then the optimal $T$ is

$$T = \sqrt{\frac{2A_D}{\lambda w}}.$$ 

Obviously, for our problem

$$\bar{Q} = \sqrt{\frac{2A_R\lambda}{h}}, \quad T = \sqrt{\frac{2A_D}{\lambda w}},$$

is a suboptimal solution. As we illustrate in the following section, Equation (41) seems to be an accurate approximation for our problem. On the other hand, the suboptimal solution in Equation (42) may perform significantly worse, especially if $h$ is large.

### 6. Numerical Illustrations

In this section we compute $(Q^*, T^*)$ and the solution in Equation (41), and substitute them into Equation (28) to compare the resulting cost values. We utilize Theorem 2 and Lemma 1 for computing $(Q^*, T^*)$. We note that, in our numerical examples, $Q^*, \bar{Q}$, and cost function values are rounded to the nearest integer.

The base values of model parameters are given as $A_R = 125$ per replenishment, $h = 7$ per unit per week, $A_D = 50$ per delivery, and $w = 10$ per unit per week, $\lambda = 10$ units per week. In Table 1, we summarize our calculations for the base parameter values. For this particular case, Lemma 1 implies that

$$18.36 \leq Q^* \leq 18.89.$$ 

Computing the solution of Equation (30) which satisfies the above inequality, we obtain that $Q^* = 18.54 \approx 19$ and $T^* = .66$. We also find that the approximate

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Solutions for Base Parameter Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{Q}$</td>
<td>$T$</td>
</tr>
<tr>
<td>Optimal Solution</td>
<td>19</td>
</tr>
<tr>
<td>Solution in Equation (41)</td>
<td>19</td>
</tr>
</tbody>
</table>
solution in Equation (41) is given by (19, .65), and it is very close to our optimal solution (19, .65), and it is very close to our optimal solution (19, .66).

In Tables 2–6 we provide additional numerical examples by varying one parameter at a time while keeping others at base values. We find that the bounds on \( Q^* \) provided by Lemma 1 are very tight for all of these numerical problems. Again, the results verify our intuitive conclusion that Equation (41) provides a satisfactory approximation. Our numerical examples also illustrate the sensitivity of the solution relative to the model parameters. For example,

- as \( A_R \) increases the resulting \( Q^* \) and cost values increase;
- as \( \lambda \) increases the corresponding \( Q^* \) increases whereas the corresponding \( T^* \) decreases;
- as \( h \) increases the resulting \( Q^* \) and \( T^* \) decreases,
- as \( A_D \) increases the corresponding \( T^* \) value increases; and
- as \( w \) increases \( T^* \) decreases.

### 7. A Cost Comparison

As we have already mentioned, under the assumptions of the classical inventory models the demands are satisfied as they arrive. That is, the traditional literature focuses on immediate delivery policies. Therefore, parameters \( A_D \) and \( c_D \) are sunk costs, and

---

**Table 2** Solutions for Varying \( A_k \) Values

<table>
<thead>
<tr>
<th>( A_k )</th>
<th>Value of Equation (28)</th>
<th>Value of Equation (40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>N/A</td>
<td>267</td>
</tr>
</tbody>
</table>

**Table 3** Solutions for Varying \( \lambda \) Values

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>Value of Equation (28)</th>
<th>Value of Equation (40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>N/A</td>
<td>197</td>
</tr>
</tbody>
</table>

**Table 4** Solutions for Varying \( h \) Values

<table>
<thead>
<tr>
<th>( h )</th>
<th>Value of Equation (28)</th>
<th>Value of Equation (40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>N/A</td>
<td>249</td>
</tr>
</tbody>
</table>

**Table 5** Solutions for Varying \( A_D \) Values

<table>
<thead>
<tr>
<th>( A_D )</th>
<th>Value of Equation (28)</th>
<th>Value of Equation (40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>237</td>
<td>238</td>
</tr>
</tbody>
</table>

**Table 6** Solutions for Varying \( w \) Values

<table>
<thead>
<tr>
<th>( w )</th>
<th>Value of Equation (28)</th>
<th>Value of Equation (40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>275</td>
<td>277</td>
</tr>
</tbody>
</table>

---
they are not modeled explicitly. In this section, we provide a point of comparison between the immediate delivery policy and the consolidated delivery policy proposed in the previous section.

If demands are delivered as they arrive (without consolidation), then no waiting costs accumulate. In this case, our problem reduces to a pure inventory problem for which the order-up-to level (denoted by $Q$ as in the previous section) and reorder level (denoted by $s$) should be computed. Since the replenishments are instantaneous, the optimal reorder level is again zero. That is, there is no need to replenish inventory unless the inventory level is zero when a demand arrives and a shipment should be made. Computing the optimal $Q$ is also straightforward. Considering the case of Poisson demand with rate $\lambda$, expected long-run average cost under an immediate delivery policy is given by (see Bhat 1984, pp. 435–436):

$$\frac{A_R \lambda}{Q} + c_R \lambda + \frac{h(Q + 1)}{2} + A_D \lambda + c_D \lambda.$$  \hspace{1cm} (43)

Thus, the optimal $Q$ value is again given by the standard EOQ formula as in Expression (41). Using Expression (43), we can compare the cost of an immediate delivery policy and a consolidated delivery policy.

For the base problem discussed in §6, the weekly cost of the optimal consolidated delivery policy is $281 (see Table 1.). On the other hand, substituting the base parameter values and $Q = 19$ into Expression (43), the weekly cost of the optimal immediate delivery policy is computed as $636. In fact, provided that $h$ and $w$ are small, and $\lambda$ and $A_D$ are large, the consolidated delivery policy outperforms the immediate delivery policy. However, for slow moving items, i.e., if $\lambda$ is sufficiently small, an immediate delivery policy may be more desirable. For example, if we assume $\lambda = 0.5$ for our base problem, then the weekly cost of the optimal consolidated delivery policy is $60$, whereas the weekly cost of the optimal immediate delivery policy is $58$.

8. Conclusion and Generalizations
This paper introduces a new class of supply-chain problems applicable in the context of VMI. We believe that the renewal theoretic model presented here provides a basis for future analytical work. An important generalization of the problem requires the demand to be modeled as a compound Poisson process. Interpreting $D_n$ as the weight of the $n^{th}$ demand where $\{D_n: n = 1, 2, \ldots\}$ represents a sequence of i.i.d. random variables, we let

$$\mathcal{N}(t) = \sum_{j=1}^{N(t)} D_j.$$  \hspace{1cm}

That is, $\mathcal{N}(t)$ represents the cumulative demand process corresponding to the renewal reward process $(X_n, D_n), n = 1, 2, \ldots$ with Poisson arrivals. Letting $U_0 = 0$ and $U_n = \sum_{j=1}^{\mathcal{N}(n)} D_j$, we define

$$\hat{N}(u) = \sup\{n: U_n \leq u\}.$$  \hspace{1cm}

In turn, $\hat{N}(u)$ is a renewal process that registers the number of demand orders placed by the time cumulative demand reaches $u$.

In this study, we compute a time-based consolidation policy along with a replenishment quantity. A related research problem involves simultaneous computation of a quantity-based consolidation policy and a replenishment quantity. Recall that a quantity-based policy is specified by a critical consolidated weight. That is, a quantity-based policy ships an accumulated load when an economic freight quantity, say $Q_e$, has accumulated. Then, a variation of our model should be developed for computing the optimal $Q$ and $Q_e$ values simultaneously.

The literature on shipment consolidation also identifies a hybrid temporal consolidation policy, called time-and-quantity policy. This policy is aimed at balancing the trade-off between pure time-based and quantity-based policies. Under a time-and-quantity policy, a dispatch decision is taken at $\min\{T(Q_e), T\}$ where $T(Q_e)$ denotes the arrival time of the $Q_e^{th}$ demand. A further generalization of our research consists of more complex problems involving hybrid shipment-consolidation policies.

Considering the case where the replenishment lead-time is not negligible, the following problems also require additional research:

• Computation of an $(s, S, T)$ policy under which a delivery is made every $T$ time units, and inventory
replenishment decisions are made following an \((s, S)\) policy.

- Computation \((s, S, Q)\) policy under which a dispatch decision is taken when the consolidated weight exceeds \(Q\), and inventory replenishment decisions are made following a general \((s, S)\) policy.\(^1\)

\(^1\) Research was supported by National Science Foundation Grant DMI-9908221.

**Appendix**

**Proof for Proposition 1.** Observe that function \(H(Q, T)\) represents the expected cumulative inventory held until the next replenishment, given that starting inventory is \(Q\) units and dispatch frequency is \(T\) time units. Using the renewal argument, we can write

\[
H(Q, T|N_i(T) = i) = \begin{cases} 
TQ, & \text{if } i > Q, \\
TQ + H(Q - i, T), & \text{if } i \leq Q.
\end{cases}
\]

Thus,

\[
H(Q, T) = E[H(Q, T|N_i(T) = i)] = TQ + \sum_{i=0}^{Q} H(Q - i, T)g(i).
\]

Observe that the above expression for \(H(Q, T)\) is a discrete renewal-type equation, and its unique solution is given by Equation (9) (see Tijms 1994, p. 5.). This completes the proof. □

**Proof for Proposition 2.** Given that no customers are waiting to begin with, \(W(T)\) denotes the expected cumulative waiting during a period of \(T\) time units. Thus, conditioning on the arrival time of the first demand, we have

\[
W(T) = E[W(T|S_1 = t)] = \int_0^T (T - t)dF(t) + \int_0^T W(T - t)dF(t).
\]

The above expression for \(W(T)\) is a renewal-type integral equation. Its well known solution is given by Equation (14), and this completes the proof (see Tijms 1994, p. 5.). □

**Proof for Proposition 3.** Recalling Equation (5), we have \(P(\bar{k} \geq k + 1) = G^\bar{k+1}(Q)\), and thus \(P(\bar{k} \leq k) = 1 - G^k(Q)\). Given Expression (21), we can write

\[
P(\bar{k} \leq k) = 1 - \sum_{i=0}^{\infty} \frac{(kT)e^{-\lambda T} \lambda \bar{k}}{i!}, \quad k = 1, 2, \ldots.
\]

Treating \(k\) as a continuous variable, the right hand side of the above expression is a \(Q\)-stage Erlang distribution function with parameter \(\lambda T\) and mean \((Q + 1) / (\lambda T)\). This completes the proof. □

**Proof for Proposition 4.** By definition,

\[
m_{\bar{k}}(i) = M_{\bar{k}}(i) - M_{\bar{k}}(i - 1)
\]

i.e., \(M_{\bar{k}}(i)\) is the renewal function associated with \(G(i)\). Then, it follows from Equation (6) that

\[
E[\bar{k}] = M_1(Q) + 1.
\]

If we use Equation (24) in the above and solve \(M_1(Q)\), then

\[
M_1(Q) = \frac{Q + 1}{\lambda T} - 1.
\]

In turn, Expression (44) leads to Equation (25), and this concludes the proof. □

**Proof for Theorem 1.** In order to prove this theorem, it is sufficient to show that functions \(-C_1(\bar{Q})\) and \(C_1(\bar{Q})\) intersect at most once. Note that Expression (33) yields

\[
C_1(\bar{Q}) = -\frac{A_{\lambda}}{Q} + \frac{b}{2}.
\]

Rewriting Equation (34) we have

\[
C_1(\bar{Q}) = \frac{2A_{\lambda}(2h + w) - 4A_{\lambda}h}{Q},
\]

and letting \(\psi = 2A_{\lambda}(2h + w)\) in the above expression leads to

\[
C_2(\bar{Q}) = -\frac{A_{\lambda}h}{Q + 4\psi - 4A_{\lambda}h}.
\]

Analyzing Expression (45), we conclude that \(C_2(\bar{Q})\) is increasing over the positive axis, and thus \(-C_1(\bar{Q})\) is decreasing over \([1, +\infty)\). On the other hand, analyzing Expression (46), we observe that \(C_2(\bar{Q})\) is increasing over \([1, +\infty)\). As a result, if these two functions intersect over \([1, +\infty)\), then this intersection point should be unique. □

**Proof for Theorem 2.** If Equation (39) is violated, then Equation (38) holds. As we have already discussed, if Equation (38) is satisfied, then \(C_1(\bar{Q})\) is an increasing function over \([1, +\infty)\) so that \(Q^* = 1\). Substituting 1 for \(\bar{Q}\) in Equation (31) gives \(T^* = \sqrt{2A_{\lambda}h/w}\), and this completes the proof for the second part of the theorem.

On the other hand, under Equation (39), we have \(-C_1(1) \geq C_2\) so that \(-C_1(\bar{Q})\) and \(C_2\) intersect. Theorem 1 implies that this intersection point is unique, and thus \(C_1(\bar{Q})\) has a unique stationary point over the feasible region of our problem. Also, if Equation (39) holds then Equation (37) implies that \(C'(1) = 0\), i.e., \(C(\bar{Q})\) is decreasing at 1. At the same time, Expression (32) implies that \(C(\bar{Q})\) is increasing as \(\bar{Q}\) goes to infinity. Thus, the unique solution of Equation (36) should be the global minimizer of \(C(\bar{Q})\). It follows that, under Equation (39), \((Q^*, T^*)\) is given by Equations (30) and (31). □

**Proof for Lemma 1.** Recalling Theorem 2, if Equation (39) holds, then \(Q^*\) and \(T^*\) should satisfy Equations (30) and (31). It follows that

\[
T^* = \frac{\sqrt{2A_{\lambda}d}}{\lambda w}.
\]

Using the above inequality in Equation (30) completes the proof. □
References

Accepted by Wallace J. Hopp; received January 18, 1998. This paper has been with the authors 3 months for 2 revisions.