Operations Research Letters 42 (2014) 41–47

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

On the service performance of alternative shipment consolidation policies

Sıla Çetinkaya [a,*], Fatih Mutlu [b], Bo Wei [a]

[a] Industrial and Systems Engineering Department, Texas A&M University, College Station, TX 77843-3131, United States

[b] Mechanical and Industrial Engineering Department, Qatar University, Doha, Qatar

ARTICLE INFO

Article history:
Received 10 November 2013
Accepted 11 November 2013
Available online 19 November 2013

Keywords:
Temporal shipment consolidation
Freight consolidation
Service measures

ABSTRACT

In order to compare the performances of alternative shipment consolidation policies, we develop closed-form expressions of the distribution of the maximum waiting time and the average order delay. We examine the tradeoff between the expected delivery frequency, as measured by the expected shipment consolidation cycle length, and the average order delay. The previous analytical results regarding shipment consolidation are aimed at optimizing the performances of alternative policies using cost-based criteria whereas our results are useful under service-based criteria.

1. Introduction and related literature

"Shipment consolidation (SC) is a logistics strategy that combines two or more orders or shipments so that a larger quantity can be dispatched on the same vehicle" [10]. The existing literature offers various methods and models [2,3,5,6,8–11] for determining when to release consolidated loads while documenting motivations and benefits of SC. The previously established results are aimed mainly at optimizing the performances of different SC policies using cost-based criteria. There is no existing work that provides *analytical* results comparing the policies using service-based criteria. The current paper is potentially useful for such comparison of three classes of SC policies common in the logistics practice and literature: quantity-based policy (QP), time-based policy (QP), and hybrid policy (HP) [3,9,10].

The QP is aimed at consolidating a load of q units before releasing a shipment. There are two types of TPs. Under the first, called TP1, a shipment is made every T units of time, and all orders that arrive between the two shipment epochs are consolidated. Under the other, called TP2, the arrival time of the first order after a shipment is recorded, and the next shipment is made T time units after the arrival time of the first order. Likewise, there are two types of HPs. The first is a combination of QP and TP1, called HP1, and the second is a combination of QP and TP2, called HP2. Stated formally, under HP1, the goal is to consolidate a load of size q. However, if the time since the last shipment epoch exceeds T, then a shipment decision is made. Under HP2, the goal is also to consolidate a load of size q; but, if the waiting time of the first order after the last shipment exceeds T, then a shipment decision is made.

SC policies are implemented at the cost of prolonged *order holding*, and, hence, *customer waiting* due to delayed delivery associated with the time needed to accumulate a large load. The resulting *service performances* are strongly interrelated with the corresponding waiting times which need to be considered carefully for all practical purposes. Our goal in this paper is to develop analytical results for evaluating the service performances of QP, TP1, TP2, HP1, and HP2. To this end, we call the time between two consecutive shipments as a consolidation cycle and examine two service measures. The first pertains to the complementary probability distribution of maximum waiting time (*MWT*), i.e., the waiting time of the longest held order which is the order that arrives first in a consolidation cycle. The latter takes into account the average waiting time of an order, namely average order delay (*AOD*).

Both of these service measures are of practical interest for two main reasons:

- freight forwarders, third party logistics companies, and trucking companies (i.e., carriers) are often requested to quote delivery times;
- shippers/receivers would like to know when their loads/orders will arrive.

MWT is an important indicator for the carriers before they declare quotations indicating delivery times, especially in the context of time-sensitive or time-definite delivery. Likewise, *AOD* is of interest for both the carrier and shipper especially when revenues (associated with transportation and wholesale) and inventory ownership are deferred until after the deliveries. Also, the implied *MWT* and *AOD* are critical under shipment agreements when q

* Corresponding author.
E-mail addresses: sila@tamu.edu (S. Çetinkaya), fatihmutlu@qu.edu.qa (F. Mutlu), feixianxing@tamu.edu (B. Wei).

0167-6377/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.orl.2013.11.003

and/or T are dictated/fixed as part of the agreement, i.e., the carrier prefers to set/fix q equal to the truck capacity for higher utilization and the shipper requests/fixes a promised delivery/waiting time T. In order to support such practical considerations, the current paper offers (1) closed form expressions of *MWT* and *AOD* under the different policies, (2) a comparison of the implied *MWT* and *AOD* values for fixed q and/or T values, (3) a comparison of the implied *AOD* values for a fixed expected delivery frequency, which itself represents the average service time, and, hence, has a direct impact on *AOD*.

Despite the lack of analytical work comparing different shipment consolidation policies in terms of service performance, simulation-based numerical results are available [6,9]. In [9], assuming a Poisson order arrival process where the size of each order has Gamma distribution, the authors model freight discounts for large shipment sizes, along with a fixed cost associated with each shipment. Considering QP, TP2, and HP2, the simulation results indicate that QP is superior cost-wise to the other policies. It is also reported that HP2 outperforms the other policies when *AOD* is considered.

In [6], considering a two-echelon setting, the impacts of QP, TP1 and HP1 on inventory replenishment decisions are investigated. The upper echelon is a vendor serving a group of downstream locations. The vendor's problem is to optimally schedule the upstream replenishments as well as the outbound shipments to the downstream. Detailed numerical results, comparing the cost performances of TP1 and QP in outbound shipment consolidation under a Poisson order arrival process, indicate that in the long run QP always leads to lower expected costs. For the case where the vendor keeps no inventory and acts as a break-bulk terminal, it is proved that the cost-wise QP outperforms TP1. Further simulation results reveal that HP1 is not only superior cost-wise to TP1; but, it is also superior to both QP, TP1 in terms of long-run average of the expected cumulative waiting time.

The impact of consolidation has also been investigated in the context of batch (bulk) service queues [1,7,12]. This line of work focuses on the waiting time distribution for queues in equilibrium while considering a stochastic service time. Consolidation policies similar to the ones we consider are also studied in [13] but the focus is on deriving the distribution of the queue length.

2. Main results

We proceed with deriving analytical expressions quantifying the two service measures for the five policies of interest. In doing so, we assume that "under TBP-I and HP-I, even if no order arrives in the first T time units after a shipment, then the consolidation cycle clock starts over", i.e., "empty dispatches are allowed". Similar results relaxing this assumption can be obtained in a straightforward fashion as demonstrated in the working paper version (available through the corresponding author).

In keeping with the existing literature, we model the order arrivals as a Poisson process with parameter λ where X_i, $i = 1, 2, \ldots$, denotes the interarrival time of the ith order. Obviously, X_i is an exponential random variable with mean $1/\lambda$ and $S_i \equiv \sum_1^i X_n$, $i = 1, 2, \ldots$ is an Erlang random variable with parameters i and λ. Letting $F_{i,\lambda}(t)$ and $f_{i,\lambda}(t)$ denote the corresponding Erlang distribution and density, respectively, we have

$$F_{i,\lambda}(t) = 1 - \sum_{n=0}^{i-1} \frac{(\lambda t)^n e^{-\lambda t}}{n!} \quad \text{and} \quad f_{i,\lambda}(t) = \frac{\lambda(\lambda t)^{i-1} e^{-\lambda t}}{(i-1)!}. \qquad (1)$$

Also, letting $N(T)$ denote the number of orders within T units of time and defining $P_n \equiv P(N(T) = n)$, we have

$$P_n = \frac{(\lambda T)^n e^{-\lambda T}}{n!}.$$

Table 1
Summary of the complementary cumulative functions of *MWT*.

$$\bar{G}_{QP}(t) = \sum_{i=0}^{q-2} \frac{(\lambda t)^i e^{-\lambda t}}{i!}$$

$$\bar{G}_{TP1}(t) = \begin{cases} 1 - e^{-\lambda(T-t)}, & \text{if } t < T \\ 0, & \text{if } t \geq T \end{cases}$$

$$\bar{G}_{TP2}(t) = \begin{cases} 1, & \text{if } t < T \\ 0, & \text{if } t \geq T \end{cases}$$

$$\bar{G}_{HP1}(t) = \begin{cases} \left(\sum_{i=0}^{q-2} \frac{(\lambda t)^i e^{-\lambda t}}{i!}\right)(1 - e^{-\lambda(T-t)}), & \text{if } t < T \\ 0, & \text{if } t \geq T \end{cases}$$

$$\bar{G}_{HP2}(t) = \begin{cases} \sum_{i=0}^{q-2} \frac{(\lambda t)^i e^{-\lambda t}}{i!}, & \text{if } t < T \\ 0, & \text{if } t \geq T \end{cases}$$

Regardless of the policy, each shipment stochastically clears the system leading to a regenerative consolidation process with independent and identical consolidation cycles.

2.1. Maximum waiting time, MWT

The service measure *MWT* is quantified via the complementary cumulative distribution, $\bar{G}(t) \equiv P(MWT > t)$ summarized in Table 1 where each distribution is indexed by the policy type. Under QP, *MWT* is the sum of the subsequent $q-1$ interarrival times after the first order arrives, and, hence, it is Erlangian with parameters $q-1$ and λ. Under TP1, a shipment is made every T units of time. Hence, $\bar{G}_{TP1}(t) = P(0 < X_1 \leq T - t) = 1 - e^{-\lambda(T-t)}$, if $t \leq T$, and $\bar{G}_{TP1}(t) = 0$, otherwise. Under TP2, on the other hand, a shipment is made T units of time after the arrival of the first order in each consolidation cycle. HP1 is simply a combination of QP and TP1 so that $MWT > t$ if and only if $MWT > t$ for both of the corresponding QP and TP1. Likewise, HP2 is a combination of QP and TP2 and $MWT > t$ if and only if $MWT > t$ for the corresponding QP and TP2.

2.2. Average order delay, AOD

The second service measure, *AOD*, takes into account the average delay of orders before delivery. Hence, it can be obtained by applying the *Renewal Reward Theorem* (RRT), i.e.,

$$AOD = \frac{E[\text{Cumulative waiting per cycle}]}{E[\text{Number of orders arriving in a cycle}]} = \frac{E[W]}{\lambda E[C]},$$

where W denotes the sum of the waiting times of the orders within a consolidation cycle, i.e., cumulative waiting per cycle, and C denotes the length of a consolidation cycle. Again, we index *AOD*, W, and C by policy type as needed.

By associating an annual per-unit SC cost, denoted by w, for holding orders or making the customers wait, the term $wE[W]/E[C] = w\lambda AOD$ has been included in the previous cost-based models for computing the optimal policy parameters for QP, TP1, and HP2. We record AOD_{QP}, AOD_{TP1} and AOD_{HP2} expressions in Table 2 as they appear in the existing cost-based models [4–6,11] and proceed with derivations of AOD_{TP2} and AOD_{HP1} for a comparative analysis.

The derivation of AOD_{TP2} builds on the use of AOD_{TP1}. Let us mark the arrival time of the first order in a cycle and denote it by S_1. Then, a shipment is released at $S_1 + T$. Note that, since the order process is Poisson, after the arrival of the first order, the system behaves as if it is being operated under TP1 for the remaining orders in the cycle. Thus, the expected cumulative waiting time for the remaining orders is $\lambda T^2/2$, whereas the waiting time of the first order is T. It follows that $E[W_{TP2}] = T + \lambda T^2/2$. Also, it is easy to see that $E[C_{TP2}] = T + 1/\lambda$. Then, applying the RRT leads to AOD_{TP2}.

Recall that, under HP1, a shipment is released when the earliest of the following occurs: q items arrive or T units of time elapse after