Working Paper

Shipping Consolidation with Delivery Deadline and Expedited Shipment Options

Lai Wei
Stephen M. Ross School of Business
University of Michigan

Stefanus Jasin
Stephen M. Ross School of Business
University of Michigan

Roman Kapuscinski
Stephen M. Ross School of Business
University of Michigan

Ross School of Business Working Paper
Working Paper No. 1375
February 2017

Shipping Consolidation with Delivery Deadline and Expedited Shipment Options

Lai Wei, Stefanus Jasin, Roman Kapuscinski
Ross School of Business, University of Michigan
laiwi, sjasin, kapuscin@umich.edu

Problem definition: Shipment consolidation is commonly used to take advantage of the economies of scale by avoiding some of the shipping costs. However, when pending current orders are consolidated with future orders it may require more expensive expedited shipment in order to meet shorter deadlines. In this paper, we study the optimal consolidation policy focusing on the trade-off between economies of scale and expedited shipping costs. **Academic/Practical Relevance**: Our work is motivated by the prevalence of consolidation in the supply chain industry and also by its potential application for online and omni-channel retailing, especially with the rise of, so-called, on-demand logistic services. In such situations, sellers, have the flexibility to take advantage of consolidation, by deciding from which warehouse to fulfill the orders and also when to ship the orders, as long as the orders deadlines are met. **Methodology**: We use Dynamic Programming to study the optimal policy and its structure. We also conduct intensive simulation tests to show the good performance of heuristics which we proposed based on structures of the optimal policy. **Results**: The optimal policies and their structures are characterized in settings with up to two warehouses, where the impact of expedited shipment on both shipping policy and order fulfillment policy are explored. Utilizing the insights of these structural properties, two easily implementable heuristics are proposed, which perform within 1-2% of the optimal in intensive numerical tests. **Managerial Implications**: Despite the complexity of the actual optimal consolidation policy, sellers can apply the two simple heuristic policies we proposed to get near-optimal performance in various cases.

Key words: Consolidation, expedited shipment, deadline-driven logistics

1. Introduction

The total costs due to logistics usually accounts for 9% to 14% of sales of a company, depending on the industry sector and, on aggregate level, for 7.9% of the US GDP in 2015 (27th State of Logistic Report). Shipping cost alone comprises more than 60% of the total logistics costs (27th State of

Logistic Report). Thus, it is no surprise that "effectively managing shipping cost directly affects ... business' bottom line" (Fell, 2011). One commonly used strategy to save on shipping costs is shipment consolidation, i.e., combining multiple small shipments into one large one (Cetinkaya, 2005). Sellers often have a significant opportunity to take advantage of various forms of flexibility when satisfying customers' orders, choosing from *which* warehouse to fulfill the orders and also *when* to ship the orders. The latter is because there is usually a time window between the time the seller receives an order and the time by which the order must be delivered (Lee et al., 2001; Xu et al., 2009). This time window provides an opportunity for the seller to combine existing orders with new incoming orders so that several orders can be shipped together, reducing total shipping costs. However, there is a caveat: Since orders must be delivered by their guaranteed due date, delaying the shipment of some orders can potentially increase total shipping costs due to the need to use more expensive expedited shipping modes. Many logistic firms: Expedited Logistics and Freight Services, ASAP Expedited Logistics, Time Definite Services Inc, etc. increase shipping rates for faster shipping modes. Major carriers such as UPS, FedEx and USPS also offer shorter lead time deliveries (3 Day Select, and 2nd Day Air modes, Overnight Delivery etc.) for extra cost. Thus, the seller needs to carefully balance the trade-off between the benefit of consolidation and the potential increase in total costs due to the need to use expedited shipping. The key decisions faced by the seller are three-fold: (1) *When should the seller ship an order?* (2) *Which orders should be included in the shipment?* (3) *From which warehouse should the orders be shipped?* The last question is particularly important for the case where not all items are available in every warehouse, or store.

Our work is motivated by the prevalence of consolidation in the supply chain industry and also by its potential application for online and omni-channel retailing, especially with the rise of, so-called, on-demand logistic services. In the context of supply chain, for outbound logistics, companies such as Grainger Integrated Supply Operations or Intercore Group aggregate several suppliers into an integrated supplier group. Multiple orders containing different products from the same customer are handled jointly in the same supplier group, providing an opportunity for consolidation

(Narus, 1996). For inbound logistics, large retailers such as Amazon.com, can also benefit from order consolidation from multiple vendors. By using multi-stop shipment, Amazon.com can pick up orders from several vendors before transporting them to its warehouses (Cummings III, 2014).

For online retail, the frequency of orders per customer has continued to increase while the size of each order has become smaller due to the increasing popularity of subscription-based business models for delivery (e.g., Amazon Prime) and the popularity of quick-purchase button Amazon Dash (more than 200 different dash buttons were introduced and dash button orders have grown five-folds over one year, Lewis, 2006; Gil, 2014). While providing new sources of revenues, this trend intensifies the logistic pressure for online retailers. For example, Amazon.com had to increase the prime membership fee due to high shipping cost (Stone, 2014) while Jet.com and Alibris.com encourage customers to consolidate their own orders. To save shipping cost, online retailers can consolidate multiple orders placed by one customer, being aware that a faster, more expeditious, mode of transportation may be used to meet order deadlines. Similarly, in omni-channel retail, especially for those retailers that offer one-day (or several hours) delivery guarantee, orders from different customers in nearby locations can be consolidated and executed through multiple drop-offs.

In this paper, we study the optimal shipping and consolidation policy taking into account both the delivery deadlines and the availability of expedited shipping options. To gain insights on the structure of the optimal policy, we consider the setting where the seller operates up to two warehouses, or stores. To satisfy individual orders, the seller may satisfy multiple orders in one shipment. Each shipment incurs both fixed and variable costs, each of which is a function of delivery speed. We first analyze the case with only fixed cost for both one-warehouse and two-warehouse settings and, then, the case with both fixed and variable costs for one-warehouse setting. The insights from these analysis are useful for constructing near-optimal heuristic policies for the setting of two warehouses with both fixed and variable costs.

Our findings can be summarized as follows. In one-warehouse setting with only fixed cost, the optimal policy can be characterized by a sequence of time-dependent thresholds—it is optimal to

ship all pending orders in period t if the *slack time* (remaining time until the due date) of the most urgent order is smaller than or equal to a threshold τ_t. This result has an intuitive appeal and is easy to implement: Sellers can take advantage of shipment consolidation to a point where the increase in cost becomes so high that it exceeds the potential benefit of consolidation. For two-warehouse setting with only fixed cost, the optimal policy is not easy anymore to describe or analyze in general. In the simplest case, when the two warehouses are symmetric in their cost structure and order arrival probabilities, the optimal policy can be characterized by six non-linear boundaries in three-dimensional space. This highlights the non-triviality of generalizing threshold policy from one-warehouse setting to two-warehouse setting. Motivated by the simplicity of threshold policy in one-warehouse setting, we propose two heuristic policies that replace the six boundaries in two-warehouse setting with two or three constant thresholds. The first heuristic, which we call *warehouse-based* heuristic, uses two thresholds; the second heuristic *order-based* heuristic, uses three thresholds. Under warehouse-based heuristic, once the threshold for a warehouse is crossed, all orders that can be shipped from that warehouse are shipped; under order-based heuristic, once the threshold for an order *type* is crossed, all orders of that type are shipped, together with some other orders that can be consolidated. Our numerical experiments reveal that the performance of these heuristics across symmetric and asymmetric problem instances are within 1% of the optimal policy in most cases.

Adding variable cost into the model creates non-trivial subtleties in the analysis. In particular, the seller now needs to jointly decide which orders to ship and how to split orders into different packages. In one-warehouse setting, when all orders are guaranteed delivery within at most three periods, or equivalently using up to three (equally spaced) delivery modes, we show that the optimal policy is characterized by volume-dependent thresholds. For two-warehouse setting, we do not attempt to describe the structure of the optimal policy, due to already complex structure of the optimal policy with only fixed cost. Instead, given the good performance of constant thresholds heuristic policies in the fixed-cost case, we propose modified heuristic policies and show, using

intensive numerical experiments, that their average performances are within 0.29-2.31% of the optimal policy. These results provide an important insight that, despite the complexity of the actual optimal policy, simple heuristic policies perform well in most cases. We suspect that our heuristic policies can be generalized to the setting with n warehouses and consider it a future research project.

The remainder of this paper is organized as follows. Section 2 provides a brief literature review. In Sections 3 and 4, we study the case with only fixed cost in one-warehouse and two-warehouse settings, respectively. The case with both fixed and variable costs are considered in Sections 5 and 6. Finally, in Section 7, we conclude the paper and discuss potential future research directions. Proofs can be found in Appendix and the supplemental file.

2. Brief Literature Review

Two streams of literature are most closely related to our work: shipment consolidation, which studies how to combine several orders, and order fulfillment, which studies from which warehouse to fulfill the orders. The potential cost savings due to shipment consolidation have been extensively studied in the logistic literature (Daganzo, 1988; Pooley and Stenger 1992; Popken, 1994). The main trade-off considered in this literature is between constant fixed cost of shipping and inventory holding cost. Three types of consolidation policies are usually considered: (1) time-based, (2) quantity-based, and (3) hybrid, or time-and-quantity, consolidation. Time-based policy sets a pre-determined interval within which orders are accumulated and one shipment is dispatched at the end of the interval; quantity-based policy dispatches one shipment after a pre-determined quantity of orders is accumulated; and, hybrid policy releases a shipment either after a pre-determined quantity is achieved or at the end of a pre-determined time interval. Note that all these policies are heuristics – most existing consolidation literature either focuses on evaluating and comparing the performance of the three policies (Cooper, 1984; Burns et al., 1985; Campbell, 1990; Higginson and Bookbinder, 1994) or calculating the optimal parameters of these policies with or without integrating inventory decisions (Gupta and Bagchi, 1987; Axsater, 2001; Cetinkaya et al., 2000, 2008;

Popken 1994). Our work differs from the previous consolidation literature in the following three ways: (1) We derive the structure of the optimal policy instead of imposing a certain policy form; (2) we consider the case where all orders must be delivered by strict deadlines; (3) we consider the possibility of using expedited shipping modes with higher cost. Note that although expedited shipment has been considered in the inventory literature (Zhou and Chao, 2010; Caggiano, etc.,2006; Huggins and Olsen, 2003; Hoadley and Heyman, 1977), where expedited shipment is used to meet demand or reduce penalty cost, and supply chain risk management (Qi and Lee, 2015), where expedited shipment serves as a substitution to the reliability of suppliers, it has not been considered in previous consolidation literature.

The subject of order fulfillment is mainly discussed in the context of online retailing, or e-comerce, literature. The main trade-off considered in that literature is between shipping cost and future product availability since not all warehouses may stock the same products and there is typically an imbalance in the inventory level across all warehouses (Xu et al., 2009; Acimovic and Graves, 2015; Jasin and Sinha, 2015; Lei et al., 2016). However, the possibility of consolidation with future orders is not considered in this stream of literature. In our work, we allow different availability across warehouses, similar to this literature, and focus primarily on consolidation decision instead of split-fulfillment decision.

3. Single Warehouse with Only Fixed Cost

In this section, we study the case where all orders can be shipped from a single warehouse and there is a single shipping destination. This framework applies when the seller ships products to multiple customers located in geographically compact region. Multiple orders arriving in the same period are treated as a single order, even if they are for different products. We consider a finite-horizon problem with T periods. We count time backward, $t = T, \ldots, 1$, with period 1 being the last period. The probability of an order arrival in any period is $\alpha < 1$. Each incoming order must be delivered no later than d periods after its arrival. For example, an order arriving in period T must be delivered by period $T - d$. We define *slack time* s as the remaining time until the delivery

deadline, e.g., $s = 1$ means that the order must reach customer in the next period. If there are currently n pending orders that have not been delivered, then vector $\vec{s} = (s_1, s_2, \ldots, s_n)$ denotes the corresponding slack times, where $s_1 < s_2 < \cdots < s_n$. Let $F(x)$ denote the fixed cost of delivering an order in x periods. We assume that $F : \mathbf{R}^+ \to \mathbf{R}^+$ is non-increasing and convex. For convenience, we also assume $F(0) = \infty$ and $F(\infty) = 0$. That is, all delivery times are positive, while the condition $F(0) = \infty$ ensures that all orders are delivered on time. The sequence of events is as follows: At the beginning of period t, the seller observes the slack times of all pending orders $\vec{s} = (s_1, \ldots, s_n)$ and decides whether to ship some of these orders. Let $\Omega \subseteq \{1, \ldots, n\}$ be the subset of pending orders that is shipped. Since all orders must be delivered on time, the seller can choose to use an x-period delivery where $x \leq s_k$ and k is the smallest index in Ω, and incurs a shipping cost $F(x)$. After the delivery has been made (if any), a new order arrives with probability α.

3.1. Dynamic Programming (DP) Formulation

Suppose that there are currently n pending orders. While there are 2^{n-1} different ways of choosing the first shipment, it is straightforward to see that, in optimal policy, we either ship all n orders at the same time or none at all.

LEMMA 1. (*a*) *If it is optimal to ship an order at the current period, then it is optimal to ship all pending orders together.* (*b*) *The incurred shipping cost in a period is a function of the smallest slack time among the shipped orders.*

Lemma 1 allows us to consider the smallest slack time instead of the slack times of all pending orders as the state variable. Let $V_t(z)$ denote the cost-to-go function at the beginning of period t when the smallest slack time is z (i.e., $z = s_1$). We use $z = \infty$ to denote the case where there is no order. The pseudo-DP formulation for our problem is as follows.

For $t > 1$ and $1 \leq z < \infty$, we have:

$$\begin{aligned} V_t(z) &= \min \begin{cases} F(z) + V_t(\infty) & \text{Ship} \\ V_{t-1}(z-1) & \text{Do not ship} \end{cases} \\ V_t(\infty) &= \alpha V_{t-1}(d) + (1-\alpha) V_{t-1}(\infty). \end{aligned} \tag{1}$$

For $t=1$, we have: $V_1(z)=F(z)$.

All orders must be shipped by the end of the horizon, implying $V_1(z)=F(z)$. Shipping all orders in period t incurs a current cost $F(z)$ plus a future cost $V_t(\infty)$ while holding all orders to the next period reduces the slack time by one. Equation (2) corresponds to the case where there is no pending order – in such a case, either a new order arrives with probability α and the slack time becomes d, or no order arrives and the slack time remains ∞. The following proposition describes a property of $V_t(\cdot)$.

PROPOSITION 1. *$V_t(z)$ is non-increasing in $z \geq 1$ given t and is non-decreasing in t given $z \geq 1$.*

Proposition 1 has an intuitive interpretation: Smaller slack time means more urgency, which implies higher expected total shipping costs; smaller t means fewer future orders, which implies fewer future shipments and smaller expected total shipping costs.

3.2. The Optimal Policy: Its Structure and Its Properties

We now show that the optimal shipping policy has a simple threshold structure. We first state a lemma that will be used to prove this property.

LEMMA 2. *For all $t \geq 1$ and $z \geq 2$, $V_t(z-1)-V_t(z) \geq F(z-1)-F(z)$.*

Lemma 2 provides a link between the cost-to-go function and the current shipping cost. It allows us to explicitly compare the shipping cost of different alternatives.

THEOREM 1. *There exists an integer threshold τ_t such that the optimal decision in period t is to hold all pending orders if $z > \tau_t$ and to ship all of them if $z \leq \tau_t$.*

PROOF. Fix time period t. To prove the existence of a threshold τ_t, it is sufficient to show that if the optimal decision for slack time $z \geq 3$ is to ship all orders, then the optimal decision for slack time $z-1$ is also to ship all orders. (The case $z=2$ and $z=1$ are trivial because we must ship when slack time equals 1.) By DP formulation, it is optimal to ship with slack time z iff $F(z)+V_t(\infty) \leq V_{t-1}(z-1)$. By Lemma 2, $F(z-1)+V_t(\infty) = (F(z-1)-F(z))+(F(z)+V_t(\infty)) \leq$

$(F(z-2)-F(z-1))+V_{t-1}(z-1) \leq V_{t-1}(z-2)$. But, this implies that it is also optimal to ship for slack time $z-1$, which completes the proof. $\square$

Theorem 1 shows the existence of threshold τ_t for each time t. Since we assume a stationary arrival probability, using the standard convergence argument, as in the infinite-horizon literature (Gosavi, 2003), it is not difficult to show that there exists some τ^*, such that $\tau_t \to \tau^*$ as $t \to \infty$, see Figure 1. Thus, when the the seller considers a long-term shipping strategy, s/he does not need to worry about the time-dependent nature of the threshold; instead, s/he can simply use a constant threshold policy throughout all periods.

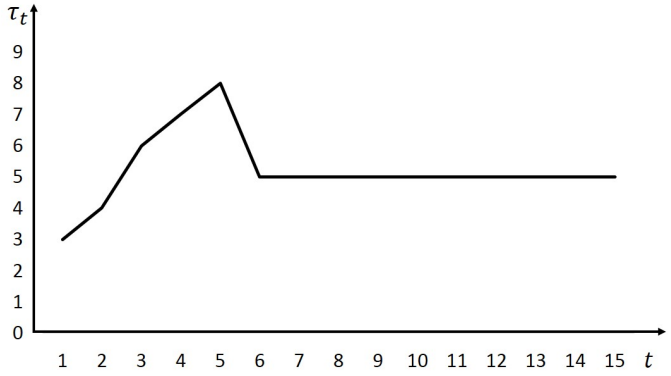

Figure 1 Threshold of slack time as a function of remaining time horizon

THEOREM 2. *Suppose that we use a constant threshold τ in all periods. Then, the expected average shipping cost during T periods converges to $C(\tau,\alpha,d)=F(\tau)\left(\frac{1}{\alpha}+d-\tau\right)^{-1}$ as $T\to\infty$.*

Theorem 2 allows us to easily compute the optimal constant threshold τ^* by minimizing $C(\tau,\alpha,d)$ over set $\{1,2,...,d\}$. One simple application of Theorem 2 is for the case where $F(\cdot)$ is linear:

PROPOSITION 2. *If the fixed cost is linear, then the optimal threshold τ is either 1 or d.*

Proposition follows because both the marginal cost and marginal benefit of waiting one additional period do not vary with slack times. For the marginal cost, this is obvious since the shipping cost is linear; as for the marginal benefit, since the arrivals are stationary, the potential saving from consolidation also does not vary with slack times.

We now switch our attention to a critical element influencing the seller's policy, the probability of order arrival α.

LEMMA 3. *The optimal constant threshold τ^* decreases as α increases.*

Lemma 3 implies that τ^* is smallest when $\alpha \approx 1$ and largest when $\alpha \approx 0$. This is intuitive: If orders arrive very frequently, the opportunity of consolidating orders is high, which provides an incentive for the seller to delay shipping; if, on the other hand, orders arrive only infrequently, the opportunity of consolidating orders is lower and it is better to ship them earlier due to the risk of incurring a high shipping cost without the benefit of consolidation.

4. Two Warehouses with Only Fixed Cost

We now consider the case where the seller operates two warehouses, W1 and W2. We classify incoming orders into three types: type A can only be fulfilled from W1, type C can only be fulfilled from W2, and type B can be fulfilled from both W1 and W2. If an order contains items of different types, we assume that the seller simply treats these as separate orders. Consequently, orders for different products may have the same due date. The optimal policy for two-warehouse setting is, in general, difficult to characterize. However, when two warehouses are symmetric, we show below that the optimal policy is characterized by six boundaries, which are functions of the slack times of all order types. While this result is a generalization of the threshold policy in one-warehouse setting, unlike the threshold policy in one-warehouse case, the six boundaries in the two-warehouse setting are not easy to compute optimally. To address this, we propose heuristic policies and show numerically that their performances are within 1-2% of the optimal policy.

4.1. Dimensionality Reduction and Dynamic Programming (DP) Formulation

In two-warehouse setting the seller needs to consider, for each order type, which subset of orders to ship and from which warehouse. Before we formulate this problem, we first introduce some notation. Let α_X denote the arrival probability of orders type $X \in \{A, B, C\}$. The fixed cost functions for shipping from W1 and W2, in x periods, are denoted by $F_1(x)$ and $F_2(x)$, respectively. $F_i(x)$, for $i \in \{1, 2\}$, is assumed to have the same properties (non-increasing and convex) as $F(x)$ in Section 3. The sequence of events is also the same as in Section 3 and all orders also must be delivered no later than d periods after their arrivals.

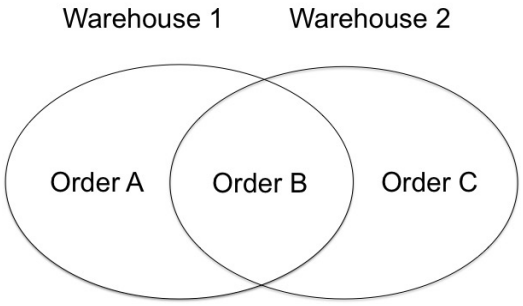

Figure 2 Problem illustration for two-warehouse setting

Similarly to one-warehouse case (Lemma 1), we show in Lemma 4 that if an order of a particular type is shipped, then all pending orders of the same type should also be shipped. Thus, the vector of the smallest slack times (z_A, z_B, z_C), for orders of types A, B, and C completely describe the state space. We refer to these simply as slack times.

LEMMA 4. (*a*) *If it is optimal to ship an order of a particular type in the current period, then it is optimal to ship all pending orders of the same type.* (*b*) *The incurred shipping cost in a period is a function of the smallest slack times among the shipped orders.*

With the simplified decision space, the pseudo-DP formulation for the two-warehouse setting is as follows. Let x_i $(i \in \{A, B, C\})$ denote the decision whether to ship orders type i; we set $x_i = 1$ if orders type i are shipped and $x_i = \infty$, otherwise. Let $V_t(z_A, z_B, z_C)$ denote the cost-to-go function at the beginning of period t. We can write $V_t(\cdot,\cdot,\cdot)$ recursively as follows:

$$\text{For } t > 1: \qquad V_t(z_A, z_B, z_C) = \min_{(x_A, x_B, x_C)} \{f(z_A x_A, z_B x_B, z_C x_C) + \mathbf{E}[V_{t-1}(\tilde{z}_A, \tilde{z}_B, \tilde{z}_C)]\}$$

$$\text{For } t = 1: \qquad V_1(z_A, z_B, z_C) = f(z_A, z_B, z_C)$$

$$\text{where} \quad f(y_1, y_2, y_3) = \min\{F_1(\min\{y_1, y_2\}) + F_2(y_3), F_1(y_1) + F_2(\min\{y_2, y_3\})\}, \quad \forall y_1, y_2, y_3$$

To make sure that all pending orders of type X are shipped when $z_X = 1$, we impose boundary conditions $V_t(0,\cdot,\cdot) = V_t(\cdot,0,\cdot) = V_t(\cdot,\cdot,0) = \infty$. $\tilde{z}$ are the new slack times, resulting from shipping decisions x_i: (1) for $z_i < \infty$, if $x_i = \infty$, then $\tilde{z}_i = z_i - 1$; if $x_i = 1$, with probability α_i, $\tilde{z}_i = d$ and, with probability $1 - \alpha_i$, $\tilde{z}_i = \infty$. (2) for $z_i = \infty$, with probability α_i, $\tilde{z}_i = d$ and, with probability $1 - \alpha_i$, $\tilde{z}_i = \infty$ $(i \in \{A, B, C\})$. Similarly to Proposition 1 in Section 3, we have:

PROPOSITION 3. *Suppose that* $1 \le z'_A \le z_A, 1 \le z'_B \le z_B$, *and* $1 \le z'_C \le z_C$. *For all* $t \ge 1$, *we have:* $V_t(z'_A, z'_B, z'_C) \ge V_t(z_A, z_B, z_C)$ *and* $V_t(z_A, z_B, z_C) \ge V_{t-1}(z_A, z_B, z_C)$.

4.2. The Optimal Policy: Its Structure and Properties

We describe the optimal policy for the case where two warehouses are symmetric, i.e., $F_1(\cdot) = F_2(\cdot)$ and $\alpha_A = \alpha_C$. We, then, briefly discuss the general case. We start by stating two lemmas that are useful for describing the general problem.

LEMMA 5. *Suppose that* $z_A, z_B < \infty$. *If it is optimal to ship orders type B from W1, then it is also optimal to ship orders type A from W1. By symmetry, if* $z_B, z_C < \infty$ *and it is optimal to ship orders type B from W2, then it is also optimal to ship orders type C from W2.*

LEMMA 6. *For all* $t \ge 1$ *and* $z_A, z_B, z_C \ge 2$, *the following holds:*

1. *If* $z_A \le z_B$, *then* $V_t(z_A - 1, z_B, z_C) - V_t(z_A, z_B, z_C) \ge F_1(z_A - 1) - F_1(z_A)$
2. *If* $z_C \le z_B$, *then* $V_t(z_A, z_B, z_C - 1) - V_t(z_A, z_B, z_C) \ge F_2(z_C - 1) - F_2(z_C)$
3. *If* $z_B \le \min\{z_A, z_C\}$, *then* $V_t(z_A, z_B - 1, z_C) - V_t(z_A, z_B, z_C) \ge \min\{F_1(z_B - 1) - F_1(z_B), F_2(z_B - 1) - F_2(z_B)\}$

Lemma 5 further simplifies the shipping alternatives by eliminating the possibility of shipping B alone from either $W1$ or $W2$ and Lemma 6 is the analog of Lemma 2 in Section 3.

The formal definition of the optimal policy and the corresponding six boundaries are given below.

LEMMA 7. *In the symmetric two-warehouse setting, for* $z_A, z_B, z_C \ge 1$, *there exist six boundaries* $\tau^{AB}_{A,t}(z_B, z_C) \le \tau^{A}_{A,t}(z_B, z_C)$, $\tau^{BC}_{C,t}(z_A, z_B) \le \tau^{C}_{C,t}(z_A, z_B)$, $\tau^{1}_{B,t}(z_A, z_C)$ *and* $\tau^{2}_{B,t}(z_A, z_C)$ *such that*

1. *If* $z_A \le \tau^{AB}_{A,t}(z_B, z_C)$, *it is optimal to ship both orders type A and B from W1; if* $z_A \le \tau^{A}_{A,t}(z_B, z_C)$, *it is optimal to ship orders type A from W1;*
2. *If* $z_B \le \tau^{1}_{B,t}(z_A, z_C)$, *it is optimal to ship both orders type A and B from W1; if* $z_B \le \tau^{2}_{B,t}(z_A, z_C)$, *it is optimal to ship orders type B and C from W2;*
3. *If* $z_C \le \tau^{BC}_{C,t}(z_A, z_B)$, *it is optimal to ship both orders type B and C from W2; if* $z_C \le \tau^{C}_{C,t}(z_A, z_B)$, *it is optimal to ship orders type C from W2.*

Moreover, the following also hold: $\tau^{AB}_{A,t}(\infty, z_C) = \tau^{A}_{A,t}(\infty, z_C)$ *and* $\tau^{BC}_{C,t}(z_A, \infty) = \tau^{C}_{C,t}(z_A, \infty)$.

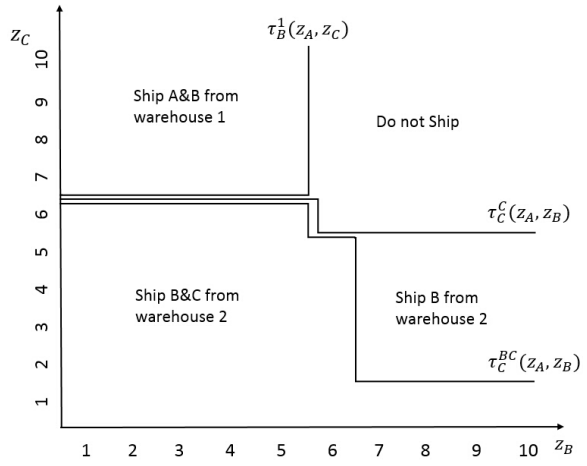

Figure 3 **Policy structure when** $z_A = 7$

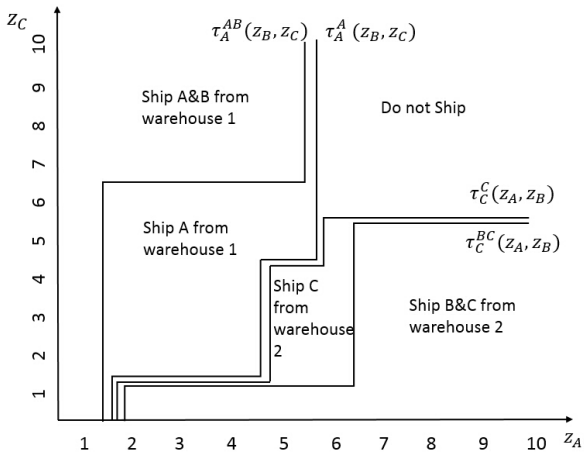

Figure 4 **Policy structure when** $z_B = 6$

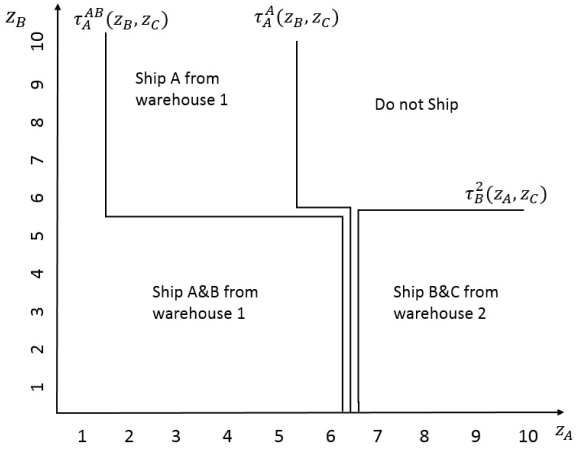

Figure 5 **Policy structure when** $z_C = 6$

Note that each of the boundaries is a function of two slack variables, so they can be viewed as surfaces in the three-dimensional space. These boundaries completely characterize the optimal shipping policy. Figures 3-5 provide illustrations of the boundaries when one of the slack variables z_A, z_B and z_C is fixed (the parameters used in this and following simulations are shown in Appendix A). Note that either $z_A \leq \tau_{A,t}^{AB}(z_B, z_C)$ or $z_A \leq \tau_{A,t}^{A}(z_B, z_C)$ implies orders type A must be shipped from W1. The reverse is also true: If it is optimal to ship orders type A from W1 in period t (regardless of whether it is also optimal to ship orders type B from W1), we must have either $z_A \leq \tau_{A,t}^{AB}(z_B, z_C)$ or $z_A \leq \tau_{A,t}^{A}(z_B, z_C)$. We state this formally below.

THEOREM 3. *The six boundaries in Lemma 7 completely characterize an optimal policy in the symmetric two-warehouse setting.*

In some cases, the six boundaries in two-warehouse setting can be reduced to thresholds, similar to these in one-warehouse setting. In an extreme case, if $\alpha_B = 0$, since type-B orders do not exist, then W1 is independent of W2 and the optimal policy for orders types A and C are each characterized by a time-dependent threshold policy. This can also be observed based on Lemma 7. Since $\alpha_B = 0$, the slack time of orders type B always equal ∞. As $\tau^{AB}_{A,t}(\infty, z_C) = \tau^{A}_{A,t}(\infty, z_C)$ and $\tau^{BC}_{C,t}(z_A, \infty) = \tau^{C}_{C,t}(z_A, \infty)$, the six boundaries reduce to only two boundaries. Further, it can be shown that these two boundaries are time-dependent constants, which is consistent with our result in Section 3. In general, however, all six boundaries are required to properly define the optimal shipping policy.

We now consider the behavior of the optimal time-independent (stationary) boundaries. The following lemma is the analog of Lemma 3.

LEMMA 8. *The optimal stationary boundaries are all decreasing in α_X, $X \in \{A, B, C\}$.*

Two things are worth noting: First, increasing the arrival probability of order type A (or C) not only decrease its own boundaries, but also the boundaries of order type B. This is because increasing the arrival probability of order type A (or C) increases the chance of future consolidation with its own type, which gives the seller more incentive to wait longer. Such incentive can be passed down to order type B, as B can be also shipped together with order type A (or C) without incurring additional cost. Second, increasing the arrival probability of order type A (or C) can also decrease the boundary for order type C (or A). This means that if the arrival probability for the orders in one warehouse increases, the seller should wait longer for the pending orders in the other warehouse. The intuition is that increasing the arrival probability for order type A (or C) leads to a lower boundary for order type B, which provides more opportunities for orders type A to be jointly shipped with orders type B without incurring additional cost.

In the asymmetric case, where either the cost functions of the two warehouses or the arrival probability of order types A and C are not the same, the optimal policy can no longer be characterized by the six boundaries in Lemma 7 (see Figures 6 and 7 for illustrations). In Figure 6, for a fixed z_A, when z_B decreases, the optimal solution for state (z_A, z_B, z_C) can change from "Do not ship" to "ship A and B from warehouse 1" and then to "Do not ship" again. In Figure 7, for a fixed z_C, the region of "Ship A from warehouse 1" is not even connected. Since the optimal policy for asymmetric case can be very complex, we do not study its structural properties; instead, we propose simple heuristic policies that can perform well for most cases. We discuss them next.

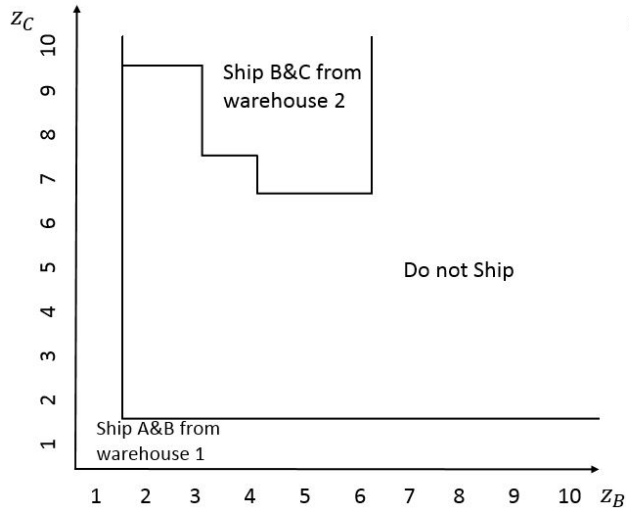

Figure 6 **Asymmetric case 1 when** $z_A = 8$

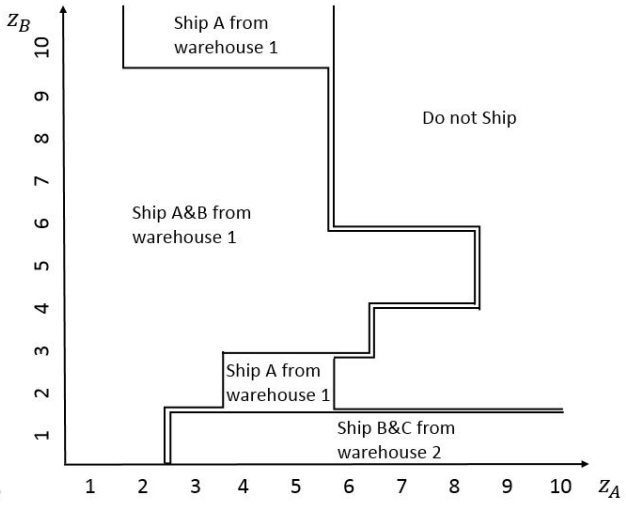

Figure 7 **Asymmetric case 2 when** $z_C = 9$

4.3. Simple Heuristic Policies

We now propose two heuristic policies that replace the six boundaries in Lemma 7 with no more than three constant thresholds for *warehouse-based* and for *order-based* heuristics. The first heuristic utilizes two thresholds, one for each warehouse, and the second heuristic utilizes three thresholds, one for each order type. Their performance is evaluated using percentage gap, defined as $(C_H - C^*)/C^*$, where C_H is the expected average cost per period by applying the heuristic and C^* is the expected average cost under the optimal policy. Their performances are tested numerically on 100 cases with a variety of costs structures and arrival probabilities of order types and they both turn out to perform very well.

4.3.1. Warehouse-Based Heuristic. Warehouse-based heuristic assigns a constant threshold for each warehouse (τ_1 for $W1$ and τ_2 for $W2$). The slack time of a warehouse is defined as the smallest slack time of orders that can be fulfilled from that warehouse. If the slack time of an order type falls below the threshold for the corresponding warehouse, then all pending orders for that warehouse are shipped. Note that this heuristic is not equivalent to treating the two warehouses independently as it allows orders type B to be fulfilled from either warehouses. The formal description of the heuristic follows:

Warehouse-Based Heuristic

Given (τ_1, τ_2) and (z_A, z_B, z_C), do:

1. If $\min\{z_A, z_B\} \leq \tau_1$, ship orders types A and B from W1 and update $z_A = z_B = \infty$;
2. If $\min\{z_B, z_C\} \leq \tau_2$, ship orders types B and C from W2 and update $z_B = z_C = \infty$.

Warehouse-based heuristic is easy to implement. It simplifies the decisions by bundling the shipment of orders types A and B (or C and B) together and reducing the number of variables to keep track of, i.e., only one slack time for each warehouse. Despite its simplicity, our numerical tests show that warehouse-based heuristic has a very good performance with the average percentage gap equal to 1.93% and its standard deviation equal to 2.39%. This suggests that the simple two-threshold warehouse-based heuristic can capture some of the key structures of the optimal six boundaries.

4.3.2. Order-Based Heuristic. Order-based heuristic assigns a constant threshold for each order type (τ_X for order type $X \in \{A, B, C\}$). It replaces the two boundaries for each order type in the optimal policy with one constant threshold: If the threshold of a particular order type is triggered, all pending orders of this type are shipped and orders of other types may also be jointly shipped according to a pre-specified consolidation rule. Basically, for each order type, the heuristic replaces the two boundaries in the optimal policy with one constant threshold and uses a myopic consolidation rule to decide whether to ship the order type alone, or with other orders from the

same warehouse. Order-based heuristic is more flexible than warehouse-based heuristic as it allows not only shipping order types A and B (or B and C) together as in the case of warehouse-based heuristic, but also shipping order type A or C alone. That said, order-based heuristic is also more complex. Note that, in warehouse-based heuristic, there is no need to decide which order types should be consolidated as it naturally consolidates all orders in the same warehouse. In order-based heuristic, when the threshold of an order type is triggered, the seller needs to properly decide whether to consolidate it with other order types.

Before we discuss the details of our proposed consolidation rule, we first highlight the importance of having a good consolidation rule when implementing order-based heuristic. Let us consider a naive threshold policy where we set a constant threshold τ_X for each order type X ($X \in \{A, B, C\}$) and implement no consolidation at all (i.e., if τ_X is the trigger, then we ship only orders type X from the cheapest warehouse). Our numerical tests reveal that the performance of this policy is very sensitive with the magnitude of α_B: For the policy to perform sufficiently well, α_B must be sufficiently close to 0; otherwise, its performance can be very poor (it can incur more than 35% of optimal cost as shown in Figure 8).

The importance of good consolidation rule is easy to explain. Consider a case where both orders type A and B are triggered and shipped jointly. With no, or poor, consolidation policy, an additional shipping cost would be incurred.

We now discuss a *one-period myopic* consolidation rule. Under our proposed rule, consolidation is decided by comparing to the best alternative, as if all orders had to be shipped in the current period. The details are shown below.

Order-Based Heuristic

Given (τ_A, τ_B, τ_C) and (z_A, z_B, z_C), do:

1. If $z_A \leq \tau_A$ and $F_1(\min\{z_A, z_B\}) + F_2(z_C) \leq F_1(z_A) + F_2(\min\{z_B, z_C\})$, ship all orders types A and B from W1; otherwise, ship only all orders type A (no consolidation);

2. If $z_B \leq \tau_B$ and $F_1(\min\{z_A, z_B\}) + F_2(z_C) \leq F_1(z_A) + F_2(\min\{z_B, z_C\})$, ship orders types A and B from W1; otherwise, ship all orders types B and C from W2;

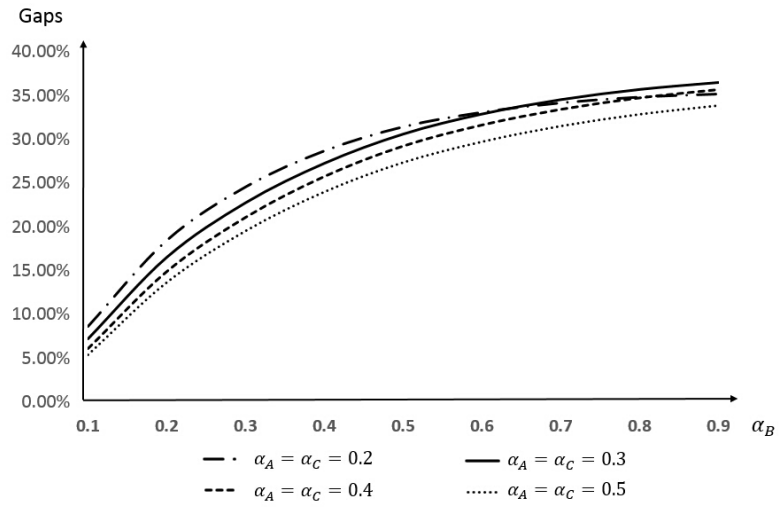

Figure 8 Gaps of Naive Order-based Heuristic

3. If $z_C \leq \tau_C$ and $F_1(z_A) + F_2(\min\{z_B, z_C\}) \leq F_1(\min\{z_A, z_B\}) + F_2(z_C)$, ship all orders types B and C from W2; otherwise, ship only all orders type C (no consolidation);

4. If more than one thresholds are hit at the same time, orders type B are the highest propriety and then A and C (e.g., if τ_B and τ_C are hit, we first proceed according to #2 and then to #3).

Our numerical experiments show that order-based heuristic performs very well, with an average percentage gap of 0.11% and a standard deviation of 0.25%. Surprisingly, this is despite the fact that the one-period myopic consolidation rule ignores expected total costs for future periods.[1]

We conclude that both warehouse-based and order-based heuristic policies perform very well. Their performance suggests that the six optimal boundaries in Theorem 3 can be well-approximated by constant thresholds, either corresponding to warehouses or order types. This approximation not only provides an easy-to-implement heuristic policy, but also justifies the use of these heuristic policies in a more complex setting in Sections 6.

[1] Interestingly, the myopic consolidation rule described above, although based on one-period comparisons, performs slightly better than two-period myopic consolidation rule, where the total shipment costs are calculated as if all orders must be shipped in the next two periods.

5. One Warehouse with Fixed and Variable Costs

In this section, we study one-warehouse setting with fixed and variable costs. All orders are assumed to have the same volume of one unit, but there may be multiple orders with the same due date. Both the fixed cost $F(z)$ and variable shipping cost $v(z)$ are determined by the slack time z of the most urgent order in the package.

That is, if orders with slack time $(z_1, z_2, \ldots, z_l)$ are shipped in one package, then the shipping cost is $F(z_1) + l \cdot v(z_1)$. We assume that $v(.)$ is convex and non-increasing function, similar to the assumption we imposed for $F(.)$. While the optimal policy for the case with only fixed cost is easy to characterize (Section 3), the optimal policy for the case with both fixed and variable costs is more challenging. First, once shipped in some period, it may no longer be optimal to ship all outstanding orders in one package (i.e., package splitting is possible).[2] Second, when some orders are shipped, it may no longer be optimal to ship all pending orders in the same period. It may be more economical (cheaper) to delay shipment of some orders in order to consolidate them with future orders.[3]

Due to the monotonicity of fixed and variable costs, for a given set of orders to ship, with slack time $\vec{s} = (z_1, z_2, \ldots, z_n)$, it is easy to write cost function $C(z_1, z_2, \ldots, z_n)$ in a recursive way:

$$C(z_1, z_2, \ldots, z_n) = \min \begin{cases} F_1(z_1) + v(z_1) + C(z_2, \ldots, z_n) & \text{ship 1 order in the } 1^{st} \text{ package} \\ F_1(z_1) + 2v(z_1) + C(z_3, \ldots, z_n) & \text{ship 2 orders in the } 1^{st} \text{ package} \\ \ldots & \\ F_1(z_1) + nv(z_1) + C(\emptyset) & \text{ship } n \text{ orders in the } 1^{st} \text{ package} \end{cases}$$

[2] As an illustration, consider a case with two orders to ship, one with slack time 1 and the other with slack time 10. The costs are $F(1) = 20$, $v(1) = 10$, and $F(10) = 2$, $v(10) = 1$. The cost of shipping the orders in two separate packages is $20 + 10 + 2 + 1 = 33$, which is smaller than the cost of shipping both orders in one package $20 + 2 \times 10 = 40$.

[3] Consider the case where $d = 10$ and there are two orders to ship, with slack time 1 and 10 respectively. The costs are $F(1) = 20$, $v(1) = 10$, $F(9) = 2.5$, $v(9) = 1.25$, and $F(10) = 2$, $v(10) = 1$. Knowing that a new (third) order arrives in the next period, the cost of shipping only the order with slack time of 1 in the current period and shipping orders 2 and 3 in the next period is $20 + 10 + 2.5 + 2 \times 1.25 = 35$, which is smaller than the cost of shipping orders 1 and 2 in the current period and shipping the new order 3 in the next period $20 + 10 + 2 + 1 + 2 + 1 = 36$.

$$C(\emptyset) = 0$$

In what follows, we first discuss the shipping cost of orders and the pseudo-DP formulation. Next, we show that the optimal policy is a volume-dependent threshold policy. To simplify this policy, we approximate the optimal threshold with a constant and show, using numerical experiments, that this approximation incurs very small additional cost compared to the optimal policy.

5.1. Dynamic Programming (DP) Formulation

To calculate the shipping cost of orders, we first need to consider how to split them into shipments. Given m orders with slack times $z_1, z_2, \ldots, z_m$ ($z_1 < z_2 < \cdots < z_m$), the number of package splitting alternatives is very high ($\frac{1}{e}\sum_{k=0}^{\infty}\frac{k^m}{k!}$, according to Lovasz, 1993).

Fortunately, the optimal package-splitting policy is monotonic, implying that if two orders z_{i_1} and z_{i_2} ($i_1 \leq i_2$) are included in the same package, then all orders $i_1 \leq j \leq i_2$ must also be included in the package (Lemma 9).

LEMMA 9. *If z_i is the most urgent order in a package and n orders are to be shipped in this package, then the other orders included in the same package should be $z_{i+1}, \ldots, z_{i+n-1}$.*

However, the optimal order-shipping policy is not necessary to be monotonic. In other words, for m ($\forall m > 0$) pending orders, if it is to ships $l\,(\leq m)$ pending orders under a policy, it is not necessarily optimal to ship the most urgent l orders. For example, for state $(z_1, z_2, z_3, z_4, z_5, z_6) = (4, 8, 9, 10, 11, 12)$, shipping orders $(8, 9, 10, 11, 12)$ and leaving order (4) to future periods can be more economic comparing to shipping orders $(4, 8, 9, 10, 11)$ instead.[4]

In general, characterizing the optimal shipping policy needs to consider two joint optimization problems, which orders to ship and how to split the current-period shipment into packages. The lack of monotonicity of order-shipping policy and the lack of closed-form expression for the cost

[4] Consider the case where $d = 12$, $F(.) = 15$, $v(z) = 13 - z$ for $0 < z \leq 8$ and $v(z) = 5$ for $8 < z \leq 12$. In period t, for state $(z_1, z_2, z_3, z_4, z_5, z_6) = (4, 8, 9, 10, 11, 12)$, suppose a policy suggests to ship 5 orders in period t and the remaining order with a new-arrival order z_7 in period $t-3$. Shipping orders $(z_1, z_2, z_3, z_4, z_5)$ in period t and (z_6, z_7) in period $t-3$ incurs cost of 85, while shipping $(z_2, z_3, z_4, z_5, z_6)$ in period t and (z_1, z_7) in period $t-3$ incurs less cost, 79.

function $C(.)$ increase the difficulty of studying the properties of the optimal policy. Fortunately, we are able to derive some results for the case when $d \leq 3$.

THEOREM 4. *For $d \leq 3$, the optimal policy has the following properties:*

1. *If a state can be reached under the optimal policy, then it is optimal to either ship all pending orders or not ship any order.*

2. *Given m pending orders, the optimal policy can be captured by volume-dependent thresholds $\tau_t(m)$. If $z_1 \leq \tau_t(m)$, it is optimal to ship all pending orders; otherwise, it is optimal not to ship any order.*

3. *For all t, the threshold $\tau_t(m)$ is non-decreasing in m.*

Theorem 4 further simplifies the decision space: Instead of having to keep track of the slack time of all pending orders, the seller only needs to keep track of the smallest slack time z_1 and the volume of orders m; and, instead of having to consider how many orders to ship, the seller only needs to choose between shipping all orders and shipping no order. It is worth noting that when it is optimal to ship all pending orders, it may still be optimal to ship orders in more than one package.[5]

Although the results of Theorem 4 are only proved for the case $d \leq 3$, our numerical tests across instances where d varies from 6 to 10 show that these results continue to hold. The pseudo-DP formulation of our problem follows:

For $t > 1$,

$$V_t(z_1, z_2, \ldots, z_m) = \min \begin{cases} C(z_1, z_2, \ldots, z_m) + V_t(\infty) & \text{Ship } m \text{ orders} \\ \alpha V_{t-1}(z-1, m+1) + (1-\alpha) V_{t-1}(z-1, m) & \text{Do not ship} \end{cases}$$

$$V_t(\infty) = \alpha V_{t-1}(d) + (1-\alpha) V_{t-1}(\infty).$$

For $t = 1$, $V_1(z_1, z_2, \ldots, z_m) = C(z_1, z_2, \ldots, z_m)$.

[5] Consider the case where the orders have two possible volume, 1 or 100, with arrival probability α_1 and α_2 respectively. The deadline of orders is $d = 2$. The cost function is $F(1) = 100$, $F(2) = 99$, $v(1) = 2$, $v(2) = 1$. In any period, for state where there are two orders, one with volume 1 and slack time 1, the other with volume 100 and slack time 2, it is optimal to ship both orders but in two separate packages.

5.2. Heuristic Policy and Numerical Experiments

While the optimal policy in Theorem 4 can be characterized by volume-dependent thresholds, it is still not easy to implement, especially since the package splitting decision requires another layer of optimization. Motivated by the constant-threshold policy in Section 3, we propose to (1) replace the volume-dependent threshold $\tau_t(m)$ with constant τ independent of vector of orders and (2) only allow orders to be shipped in one package without ever splitting the packages. The heuristic is formally defined as follows:

Constant-Threshold One-Package Heuristic

Given τ and $(z_1, z_2, \ldots, z_n)$, do:

1. If $z_1 \leq \tau$, ship all n pending orders in one package incurring cost $F(z_1) + nv(z_1)$.
2. Otherwise, wait for future orders.

The following theorem provides a theoretical performance bound for our heuristic policy.

THEOREM 5. *Suppose that $v(\cdot) = \gamma F(\cdot)$ for some $\gamma \geq 0$. Let $\tilde{C}$ denote the average cost per period for a problem with T periods under the Constant-Threshold One-Package Heuristic and let C^* denote the optimal average cost per period. Then,*

$$\frac{\tilde{C} - C^*}{C^*} \leq \min\left\{\frac{\gamma(d-\tau)\alpha}{1+\gamma}, \frac{k}{\gamma+1-k}\right\}, \quad \textit{where } k = \tfrac{d\alpha}{1+d\alpha}.$$

Note that the bound in Theorem 5 converges to 0 if $\gamma \to 0$ or $\gamma \to 1$. This is quite intuitive, when $\gamma \approx 0$, fixed cost dominates variable cost and the optimal policy is a threshold policy as in Section 3; when $\gamma \approx 1$, variable cost dominates fixed cost and the optimal policy is to ship immediately upon order arrival (in other words, $\tau = d$).

We conduct numerical experiments to test the performance of our heuristic policy in general case. Three heuristic policies are tested: (1) constant-threshold heuristic where the thresholds $\tau(m)$ are approximated by a constant and package splitting is allowed, (2) one-package heuristic where all orders must be shipped in one package and the thresholds $\tau(m)$ can vary with order volume,

and (3) constant-threshold one-package heuristic described above. Numerical tests are conducted for 630 cases where d varies from 4 to 10. The percentage gap, compared to the optimal cost, are shown in Table 1. Note that allowing either volume-dependent thresholds or package-splitting does not significantly improve the performance of the proposed heuristic (3).

Heuristic	(1)	(2)	(3)
Varying Thresholds	No	Yes	No
Splitting-package	Yes	No	No
Average	0.24%	0.08%	0.24%
Minimum	0.00%	0.00%	0.00%
10th percentile	0.00%	0.00%	0.00%
25th percentile	0.00%	0.00%	0.00%
50th percentile	0.10%	0.00%	0.10%
75th percentile	0.39%	0.07%	0.39%
90th percentile	0.74%	0.30%	0.74%
Maximum	1.24%	0.97%	1.24%

Table 1 Percentage Gaps to Optimal Cost

Further, we conduct robustness test of heuristic (3) in cases where $F(.)$ and $v(.)$ are in different function forms. Table 2 shows the average percentage gaps of each case. The results further support the good performance of heuristic (3).

$F(.)$ \ $v(.)$	1/x	$1/x^2$	$1/\log(x+1)$	e^{-x}	quadratic
$1/x$	-	0.00%	0.00%	0.01%	0.00%
$1/x^2$	0.00%	-	0.00%	0.00%	0.00%
$1/\log(x+1)$	0.07%	0.36%	-	0.23%	0.00%
e^{-x}	0.00%	0.00%	0.00%	-	0.00%
quadratic	0.56%	0.58%	0.28%	0.52%	-

Table 2 Percentage Gaps with Different Cost Functions

6. Two Warehouses with Fixed and Variable Costs

In this section, we consider two-warehouse setting with both fixed and variable costs. Note that the same challenges discussed in Section 5 continue to appear in this case, with a new complicating factor: not only a joint optimization of which orders to ship and how to split orders need to be solved, but also the possible consolidation across different order types need to be considered. Given the already complex structure of the optimal policy for two-warehouse setting with only fixed cost, we do not attempt to characterize the optimal policy. Instead, we propose and numerically test heuristic policies inspired by our results in the previous sections. In particular, since constant-threshold heuristic policies perform very well in the case with only fixed cost, we propose two constant-threshold heuristics with some modifications to account for the variable cost.

6.1. Proposed Heuristic Policies

We propose *Warehouse-Based One-Package Heuristic* and *Order-Based One-Package Heuristic*. The first heuristic is exactly the same as warehouse-based heuristic in Section 4. The second heuristic is very similar to order-based heuristic in Section 4, with only a minor difference: In Section 4, as only fixed cost in considered, consolidation is decided by the comparison between $F_1(\min\{z_A, z_B\}) + F_2(z_C)$ and $F_1(z_A) + F_2(\min\{z_B, z_C\})$. To incorporate variable cost, consolidation is now decided by the comparison between $C_1(\min\{z_{A,1}, z_{B,1}\}, n_A + n_B) + C_2(z_C, n_C)$ and $C_1(z_{A,1}, n_A) + C_2(\min\{z_{B,1}, z_{C,1}\},\ n_B + n_C)$, where $z_{X,1}$ $(X \in \{A, B, C\})$ denote the smallest slack time for order type X and $C_i(z, n) = F_i(z) + n \cdot v_i(z)$ $(i \in \{1, 2\})$. For brevity, in the following discussions we refer to these two heuristics simply as warehouse-based and order-based heuristics.

6.2. Comparison to Other Commonly Used Heuristic Policies

We test the performance of warehouse-based and order-based heuristic policies in a large scale simulation experiments with 1,458 different problem instances. We compare the performance of our proposed heuristic policies with three commonly used heuristic policies (1) Myopic heuristic, which ships orders immediately upon arrival; (2) Time-threshold heuristic, which ships orders every several periods, from each of the warehouses; and (3) Volume-threshold heuristic, which ships all

pending orders of a certain type whenever the volume of that type triggers a threshold. Myopic policies are simple and popular, e.g., they are implemented in Amazon.com (Ng 2012). Time- and volume-threshold heuristics, as discussed in Section 2, are the most widely considered heuristic policies in the consolidation literature (Cooper, 1984; Higginson and Bookbinder, 1994, Gupta and Cetinkaya et al., 2000, 2008). All three heuristics are appropriately modified to allow consolidation across the warehouses.

We separately describe the results for small and large size problems. For small-size problem ($d \leq 5$), the optimal DP policy can be numerically solved and, thus, the costs of all heuristic policies are compared with the optimal cost, see Table 3. Note that both warehouse-based and order-based heuristic policies are close to optimal and they clearly outperform the other benchmark heuristic policies by significant margins.

d	Warehouse-Based	Order-Based	Myopic	Time-threshold	Volume-threshold
3	0.66%	0.11%	15.44%	15.44%	1.22%
4	1.60%	0.29%	20.16%	6.99%	3.32%
5	2.44%	1.43%	25.83%	6.58%	7.98%

Table 3 Percentage gap to Optimal

For large size problem ($d > 5$), the optimal policy cannot be solved in a reasonable time (> 40 hours). Therefore, we use order-based heuristic policy as our benchmark, i.e. the performance of other heuristic policies are measured by the percentage gap to the cost of order-based heuristic policy, see Table 4. Note that myopic heuristic policy performs very poorly while the popular time- and volume-threshold heuristic policies incur $3-9\%$ more cost compared to order-based heuristic. This highlights the value of using slack time-based shipping and consolidation policy, compared to the more popular time- and volume-based policies.

Similar to Section 5, for robustness test of the heuristics, we conducted numerical experiments where F_1, v_1, F_2 and v_2 have different function forms. The results are summarized in 4 cases: (1) v_1 has similar magnitude as v_2, while F_1 is larger than F_2; (2) v_1 is larger than v_2, while F_1 has similar magnitude as F_2; (3) v_1 is larger than v_2 and F_1 is also larger than F_2; (4) v_1 is smaller

d	Warehouse-Based	Myopic	Time-threshold	Volume-threshold
6	1.43%	37.59%	3.57%	7.41%
8	1.80%	60.22%	3.52%	8.52%
10	2.42%	80.79%	3.77%	9.52%

Table 4 Percentage gap to Order-Based Heuristic

than v_2, while F_1 is larger than F_2. The percentage gaps, shown in Table 5, are consistent with the heuristics performances in previous tests.

case	Warehouse-Based	Myopic	Time-threshold	Volume-threshold
1	0.15%	53.82%	18.23%	0.46%
2	2.30%	67.32%	21.99%	0.84%
3	0.48%	44.44%	10.61%	2.79%
4	0.69%	42.40%	11.26%	2.64%

Table 5 Percentage gap to Order-Based Heuristic

While warehouse-based heuristic may incur about 2% higher cost than order-based heuristic, it should be noted that warehouse-based heuristic is much simpler to implement than order-based heuristic and is easily scalable to the setting with n warehouses because of its natural by-warehouse consolidation rule. In contrast, for order-based heuristic, the exact consolidation rule must be carefully constructed.

It is also worth noting that the performance gap between the two heuristics shrinks when either (1) the two warehouses are symmetric or (2) orders arrive more frequently to the cheaper warehouse. The intuition for (1) is as follows: In general, when shipping order type A (or C), the order-based heuristics can purposefully delay type B order and ship from the cheaper warehouse later, while the warehouse-based heuristic cannot do it. Such delay, however, is less likely to save costs when the warehouse costs and order arrivals are symmetric. For (2), let warehouse 1 be the cheaper one. When order type A triggers a shipment from warehouse 1, the order-based heuristic tends to also ship order type B, as warehouse 1 has lower shipping cost. Note that this is consistent with the policy of the warehouse-based heuristic. Thus, the higher the arrival probability of order

type A, the more frequently such an outcome is achieved. Observations from the numerical tests further support such insights: For (1), the warehouse-based heuristic incurs 2.52% more cost than the order-based one in general cases, comparing to only 0.55% in symmetric case. For (2), the gap between warehouse-based and order-based is only 1.31% when $\alpha_A = 0.9$, while 2.96% when $\alpha_A < 0.9$.

7. Conclusion

We analyze a practical trade-off in consolidation of shipments that has not been addressed in literature. Specifically, combining several orders into one shipment can reduce the total shipping costs; however, waiting to consolidate current orders with some future ones may require expedited shipping, thus, increasing the costs. In this paper, we study the optimal consolidation policy, focusing on the trade-off between economies of scale (combining multiple orders) and expedited shipping costs (shorter delivery window). With only fixed cost, if all orders are shipped from the same warehouse, we show that the optimal policy can be characterized by a sequence of time-dependent thresholds. With two warehouses and overlapping availability of products, the optimal policy is, in general, complex. We show that in the simplest symmetric case, the optimal policy can be characterized by six non-linear boundaries in three-dimensional space. In two-warehouse case with asymmetric fixed costs, we show that heuristics that replace the six boundaries with no more than three constant thresholds, perform very well (within 1% of the optimal policy) in most of numerically tested cases. With both fixed cost and variable cost, the difficulty of analysis increases. In one-warehouse case, the optimal policy is shown to be characterized by volume-dependent thresholds, while in two-warehouse case, we show through numerical tests that constant-threshold heuristic policies performs within 0.29 - 2.31% of the optimal policy and significantly outperform other commonly used naive heuristics.

Appendix A: Parameters in simulation experiments

In Section 4.2: In Figures 3-5: $d = 10$, $(\alpha_A, \alpha_B, \alpha_C) = (0.1, 0.1, 0.1)$, $F_1 = F_2 = F(x) = ax^2 + bx + c$, where $a = 2.5$, $b = -27.5$, $c = 130$. In Figures 6 and 7: $d = 10$, $T = 100$, $(\alpha_A, \alpha_B, \alpha_C) = (0.1, 0.1, 0.1)$, $F_1 = ax^2 + bx + c$,

where $a = 2.5$, $b = -27.5$, $c = 130$. $F_2(z) = 100 - 2z$ in Figure 6 and $F2 = 1/x + 50$ in Figure 7. **In Section 4.3**: $d = 5$. $F_1 = ax^2 + bx + c$, where $a = 2.5$, $b = -27.5$, $c = 130$. $F_2 = F_1 + \beta_1$ or $F_2 = F_1 * \beta_2$, $\beta_1 \in [0, 90]$ and $\beta_2 \in [0.01, 0.9]$. α_X ($x \in \{A, B, C\}$) varies in [0.1, 0.9]. In Figure 8, $d = 5$, $T = 20$, $F_1(x) = 1 + 3/x$ and $F_2(x) = 2 + 5/x$. **In Section 5.2**: In Table 1, variable cost is proportional to fixed cost ($v(.) = \gamma F(.)$ where $\gamma \in (0, 1)$), $\alpha \in [0.10.9]$. The linear relation between variable cost and fixed cost is a good approximation for UPS rates: fitting data for next day air, next day, 3 day, ground rate data in linear relation results in $R^2 \geq 0.8$. The rate of UPS, as one of the major courier, can be viewed as a good representative of shipping costs. In Table 2, $d = 6$, $\alpha \in (0, 1)$, and quadratic function refers to $g(x) = ax^2 + bx + c$, where $a = 2.5$, $b = -27.5$, $c = 130$. **In Section 6.2**: $d \in \{3, 4, \ldots, 10\}$, $F_1 = ax^2 + bx + c$, where $a = 2.5$, $b = -27.5$, $c = 130$, $F_2 = F_1 + \beta_1$ or $F_2 = F_1 * \beta_2$, where $\beta_1 \in [0, 90]$ and $\beta_2 \in [0.01, 0.9]$, $\alpha_X \in [0.1, 0.9]$ ($x \in \{A, B, C\}$) , and $v_i(.) = \gamma F_i(.)$, $i \in \{1, 2\}$ where $\gamma \in [0.1, 0.9]$. In Table 5, F_i is in the form of either quadratic or $1/\log(x + 1)$ and v_i either $1/x$ or $1/x^2$ (such function forms generated the largest percentage gaps in Section 5).

Appendix B: Proofs in section 3

Proof of Lemma 1: Suppose that we are at the beginning of period t and there are n pending orders that have not been fulfilled. If it is optimal to ship order 1, then it is also optimal to ship orders $2, 3, ..., n$ because including orders $2, 3, ..., n$ does not increase the current shipping cost. If, on the other hand, it is not optimal to ship order 1 in period t, then it is also not optimal to ship any subset of orders $2, 3, ..., n$ in period t. To see this, suppose that it is optimal to ship orders $S = \{i_1, i_2, ..., i_k\}$, where $1 < i_1 < i_2 < ... < i_k$. Consider the following alternative shipping policy: instead of shipping S in period t, we ship them in a later period $t' < t$ when order 1 is shipped. The current shipping cost is saved and no new additional cost is incurred, which contradicts the optimality of shipping orders in S.

Proof of Lemma 2: We first show that the cost-to-go function $V_t(z)$ can also be written as: $V_t(z) = \min\{F(z) + V_t(\infty), F(z-1) + V_{t-1}(\infty), \ldots, F(z-k) + V_{t-k}(\infty)\}$ where $k = \min\{t, z\} - 1$. The terms after the equality represent the costs of different alternatives. For example, $F(z) + V_t(\infty)$ is the cost of shipping all orders in period t, $F(z-1) + V_{t-1}(\infty)$ is the cost of delaying for one period and shipping all orders in period $t - 1$, etc. It is important to note that the cost-to-go function $V_t(z)$ is completely characterized by the values of $V_t(\infty)$ for all z. We then show a technical lemma which shows that the difference between the minimum of two set of numbers is larger than the minimal pairwise difference.

LEMMA 10. *Define* $x = \min\{a_1 + b_1, ..., a_n + b_n\}$ *and* $y = \min\{a_1, ..., a_n\}$. *If* $b_1 \geq b_2 ... \geq b_n$, *then* $x - y \geq b_n$.

Proof. Suppose that $x = a_k + b_k$ for some k. Then, $x - y \geq (a_k + b_k) - a_k = b_k \geq b_n$. $\square$

Then we show the proof of Lemma 2. Suppose that $t \geq d$ (the case $t < d$ can be proved in a similar manner and so is omitted). We can write: $V_t(z-1) = \min\{F(z-1) + V_t(\infty), F(z-2) + V_{t-1}(\infty), \cdots, F(1) + V_{t-(z-2)}\}$ and $V_t(z) = \min\{F(z) + V_t(\infty), F(z-1) + V_{t-1}(\infty), \cdots, F(1) + V_{t-(z-1)}(\infty)\} \leq \min\{F(z) + V_t(\infty), F(z-1) + V_{t-1}(\infty), \cdots, F(2) + V_{t-(z-2)}(\infty)\}$. By the convexity of $F(.)$, $F(1) - F(2) \geq F(2) - F(3) \geq ... \geq F(z-1) - F(z)$. So, by lemma 10, $V_t(z-1) - V_t(z) \geq V_t(z-1) - \tilde{V}_t(z) \geq F(z-1) - F(z)$.

Appendix C: Proofs in Section 4

Proof of Lemma 4: (a) For order type i ($i \in \{A, B, C\}$) , suppose that we are at the beginning of period t and there are n pending orders that have not been fulfilled. If it is optimal to ship order 1, then it is also optimal to ship orders $2, 3, ..., n$ because including orders $2, 3, ..., n$ does not increase the current shipping cost. If, on the other hand, it is not optimal to ship order 1 in period t, then it is also not optimal to ship any subset of orders $2, 3, ..., n$ in period t. To see this, suppose that it is optimal to ship orders $S = \{i_1, i_2, ..., i_k\}$, where $1 < i_1 < i_2 < ... < i_k$. Consider the following alternative shipping policy: instead of shipping S in period t, we ship them in a later period $t' < t$ when order 1 is shipped. The current shipping cost is saved and no new additional cost is incurred, which contradicts the optimality of shipping orders in S. (b) As the orders should be shipped using one shipment and the earliest order should meet the due date, the shipping cost is a function of the smallest slack time of orders.

Proof of Lemma 6: We prove it by induction.

For $t = 1$: for $z_A \leq z_B$, $V_t(z_A, z_B, z_C) = \min\{F_1(\min\{z_A, z_B\}) + F_2(z_C), F_1(z_A) + F_2(\min\{z_B, z_C\})\} = F_1(z_A) + F_2(z_C)$. Thus, $V_t(z_A - 1, z_B, z_C) - V_t(z_A, z_B, z_C) \geq F_1(z_A - 1) - F_1(z_A)$. Similarly, for $z_C \leq z_B$, $V_t(z_A, z_B, z_C) = F_1(z_A) + F_2(z_C) = F_1(z_A) + F_2(z_C)$ and $V_t(z_A, z_B, z_C - 1) - V_t(z_A, z_B, z_C) \geq F_2(z_C - 1) - F_2(z_C)$; For $z_B \leq \min\{z_A, z_C\}$, $V_t(z_A, z_B, z_C) = \min\{F_1(z_B) + F_2(z_C), F_1(z_A) + F_2(z_B)\}$. Thus, $V_t(z_A, z_B - 1, z_C) - V_t(z_A, z_B, z_C) \geq \min\{F_1(z_B - 1) - F_1(z_B), F_2(z_B - 1) - F_2(z_B)\}$, where the last inequality is from Lemma 10. Then, suppose the inequalities hold for $t \leq t'$ (for some t'). Then for $t = t' + 1$, we compare the cost of all the possible shipping alternative in $V_t(z_A, z_B, z_C)$ and $V_t(z_A, z_B, z_C)$.

We first prove the $V_t(z_A - 1, z_B, z_C) - V_t(z_A, z_B, z_C)$ part for $z_A \leq z_B$. We divide the shipping alternatives into the following 2 cases. **(1)** For (x_A, x_B, x_C) with $x_A = 1$, $C_{(x_A, x_B, x_C)}(z_A - 1, z_B, z_C) - C_{(x_A, x_B, x_C)}(z_A, z_B, z_C) = f(z_A - 1, z_B x_B, z_C x_C) - f(z_A, z_B x_B, z_C x_C) = F_1(z_A - 1) + F_2(z_C x_C) - F_1(z_A) - F_2(z_C x_C) = F_1(z_A - 1) - F_1(z_A)$. The first equation is from the fact that when $x_A = 1$, $\mathbf{E}[V_{t-1}(\tilde{z}_A, \tilde{z}_B, \tilde{z}_C)] =$

$\mathbf{E}[V_{t-1}(\tilde{z_A'}, \tilde{z}_B, \tilde{z}_C)]$ where $(z_A' = z_A - 1)$. And the second equation is from $z_A \leq z_B$ and simple algebra. **(2)** For $x_A = \infty$, $C_{(x_A,x_B,x_C)}(z_A - 1, z_B, z_C) - C_{(x_A,x_B,x_C)}(z_A, z_B, z_C) = \mathbf{E}[V_{t-1}(z_A - 2, \tilde{z}_B, \tilde{z}_C)] - \mathbf{E}[V_{t-1}(z_A - 1, \tilde{z}_B, \tilde{z}_C)] \geq F_1(z_A - 1) - F_1(z_A)$. The last inequality is from the linearity of expectation and induction hypothesis that for each scenario of $\tilde{z}_B$ and $\tilde{z}_C$, $V_{t-1}(z_A - 2, \tilde{z}_B, \tilde{z}_C) - V_{t-1}(z_A - 1, \tilde{z}_B, \tilde{z}_C) \geq F_1(z_A - 2) - F_1(z_A - 1) \geq F_1(z_A - 1) - F_1(z_A)$. With **(1)** and **(2)**, the proof for the $V_t(z_A - 1, z_B, z_C) - V_t(z_A, z_B, z_C)$ part for $z_A \leq z_B$ is complete. By the symmetric structure of product A and C, it is easy to see that the $V_t(z_A, z_B, z_C - 1) - V_t(z_A, z_B, z_C)$ part for $z_C \leq z_B$ can be proved by similar logic as above.

Then, we prove the $V_t(z_A, z_B - 1, z_C) - V_t(z_A, z_B, z_C)$ part for $z_B \leq \min\{z_A, z_C\}$. We divide the 8 alternatives into the following 2 cases. **(1)** For $x_B = 1$, $C_{((x_A,x_B,x_C))}(z_A, z_B - 1, z_C) - C_{((x_A,x_B,x_C))}(z_A, z_B, z_C) = f(z_A x_A, z_B - 1, z_C x_C) - f(z_A x_A, z_B, z_C x_C) \geq \min\{F_1(z_B - 1) - F_1(z_B), F_2(z_B - 1) - F_2(z_B)\}$. The first equation is from the fact that when $x_B = 1$, $\mathbf{E}[V_{t-1}(\tilde{z}_A, \tilde{z}_B, \tilde{z}_C)] = \mathbf{E}[V_{t-1}(\tilde{z}_A, \tilde{z_B'}, \tilde{z}_C)]$ where $(z_B' = z_B - 1)$. And the second inequality is from $z_B \leq \min\{z_A, z_C\}$, Lemma 10 and simple algebra. **(2)** For $x_B = \infty$, $C_{((x_A,x_B,x_C))}(z_A, z_B - 1, z_C) - C_{((x_A,x_B,x_C))}(z_A, z_B, z_C) = \mathbf{E}[V_{t-1}(\tilde{z}_A, z_B - 2, \tilde{z}_C)] - \mathbf{E}[V_{t-1}(\tilde{z}_A, z_B - 1, \tilde{z}_C)] \geq \min\{F_1(z_B - 1) - F_1(z_B), F_2(z_B - 1) - F_2(z_B)\}$. The last inequality is from the linearity of expectation and the fact that, for each scenario of $\tilde{z}_A$ and $\tilde{z}_C$, from induction hypothesis, $V_{t-1}(\tilde{z}_A, z_B - 2, \tilde{z}_C) - V_{t-1}(\tilde{z}_A, z_B - 1, \tilde{z}_C) \geq \min\{F_1(z_B - 2) - F_1(z_B - 1), F_2(z_B - 2) - F_2(z_B - 1)\} \geq \min\{F_1(z_B - 1) - F_1(z_B), F_2(z_B - 1) - F_2(z_B)\}$. With **(1)** and **(2)**, the $z_B \leq \min\{z_A, z_C\}$ part is complete.

Appendix D: Proofs in Section 5

Proof of Lemma 9: Suppose a package (denote P_i) includes z_i as the most urgent order and n orders are shipped in it, suppose it is optimal to include $z_k (k > i + n - 1)$ in it. Then there must be some order z_j $(i < j \leq i + n - 1)$ shipped in other package (denote P_j). By exchanging z_k and z_j, the shipping cost of package P_j is reduced while the cost of package P_i doest not change. This contradicts with the optimality.

References

[1] Acimovic, Jason, and Stephen C. Graves. 2014. Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing & Service Operations Management* **17** (1) 34-51.

[2] Burns, L.D., Hall, R.W., Blumenfeld, D.E., Daganzo, C.F., 1985. Distribution strategies that minimize transportation and inventory costs. *Operations Research* **33** (3) 469-490.

[3] Caggiano, K.E., Muckstadt, J.A, Rappold, J.A, 2006. Integrated real-time capacity and inventory allocation for reparable service parts in a two-echelon supply system. *Manufacturing & Service Operations Management* **8** (3) 292-319.

[4] Campbell, J.F., 1990. Designing logistics systems by analyzing transportation, inventory and terminal cost tradeoffs. *Journal of Business Logistics* **11** (1) 159-179.

[5] Cetinkaya, S., 2005. Coordination of inventory and shipment consolidation decisions: A review of premises, models, and justification. Applications of supply chain management and e-commerce research. Springer.

[6] Cetinkaya, Sila, Tekin, E., Lee, C.-Y., 2008. A stochastic model for integrated inventory replenishment and outbound shipment release decisions. *IIE Transactions* **40** (8) 324-340.

[7] Cooper, M.C., 1984. Cost and delivery time implications of freight consolidation and warehouse strategies.*International Journal of Physical Distribution and Materials Management* **14** (6) 47-67.

[8] Council of Supply Chain Management Professionals. 2016. 27th Annual State of Logistics Report.

[9] Cummings III, Charles R. 2014. Improving the inbound supply chain through dynamic pickup windows. Diss. Massachusetts Institute of Technology.

[10] Daganzo, C.F., 1988. Shipment composition enhancement at a consolidation center. *Transportation Research Part B* **22** (2) 103-124.

[11] Fell, Jason. 2012. Five Tips for Saving Money on Shipping. Entrepreneur.com

[12] Gil, Ricard, Evsen Korkmaz, and Ozge Sahin. 2014. Optimal Pricing of Access and Secondary Goods with Repeat Purchases: Evidence from Online Grocery Shopping and Delivery Fees.

[13] Gosavi, Abhijit. 2003. Simulation-based optimization. Parametric Optimization Techniques and Reinforcement Learning. Kluwer Academic Publishers.

[14] Gupta, Y. P. 1990. A feasibility study of JIT purchasing implementation in a manufacturing facility. *International Journal of Operations and Production Management.* **10** 31-41.

[15] Gupta, Y.P., Bagchi, P.K., 1987. Inbound freight consolidation under just-in-time procurement: application of clearing models. *Journal of Business Logistics* **8** (2) 74-94.

[16] Higginson, James K., and James H. Bookbinder. 1994. Policy recommendations for a shipment consolidation program. *Journal of Business Logistics.* **15** (1) 87-112.

[17] Hoadley, Bruce, and Daniel P. Heyman. 1977. A twoechelon inventory model with purchases, dispositions, shipments, returns and transshipments. *Naval Research Logistics Quarterly* **24** (1) 1-19.

[18] Huggins, E.L., Olsen, T.L,. 2003. Supply Chain Management with Guaranteed Delivery. *Management Science* **40** (9) 1154-1167.

[19] Jaruphongsa, W., S. Cetinkaya, and C. Lee. 2005. A dynamic lot-sizing model with multi-mode replenishments: polynomial algorithms for special cases with dual and multiple modes. *IIE transactions* **37** (5) 453-467.

[20] Jasin, Stefanus, and Amitabh Sinha. 2014. LP-Based Artificial Dependency for Probabilistic Etail Order Fulfillment. *Ross School of Business Working Paper.* 1250.

[21] Lee, Chung-Yee, Sila etinkaya, and Albert PM Wagelmans. 2001. A dynamic lot-sizing model with demand time windows. *Management Science* **47** (10) 1384-1395.

[22] Lei Yanzhe, Stefanus Jasin and Amitabh Sinha. 2016. Dynamic Joint Pricing and Order Fulfillment for E-Commerce Retailers. Working paper.

[23] Lewis, M., Singh, V. and Fay, S. 2006. An empirical study of the impact of nonlinear shipping and handling fees on purchase incidence and expenditure decisions. *Marketing Science* **25** (1): 51-64.

[24] Lovasz, Laszlo. *Combinatorial problems and exercises*. Vol. 361. American Mathematical Soc., 1993.

[25] Ng, Chong Keat. Inbound supply chain optimization and process improvement. Diss. Massachusetts Institute of Technology, 2012.

[26] Pooley, John, and Alan J. Stenger. 1992. Modeling and evaluating shipment consolidation in a logistics system. *Journal of Business Logistics* **13** (2) 153.

[27] Popken, Douglas A. 1994. An algorithm for the multiattribute, multicommodity flow problem with freight consolidation and inventory costs. *Operations research* **42** (2) 274-286.

[28] Qi, Lian, and Kangbok Lee. Supply chain risk mitigations with expedited shipping. *Omega* 57 : 98-113.

[29] Stone, B., J. Brustein. 2014. As It Warned, Amazon Boosts the Price of Prime. *Bloomberg.com*

[30] Xu, Ping Josephine, Russell Allgor, and Stephen C. Graves. 2009. Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing & Service Operations Management* **11** (2) 340-355.

[31] Zhou, Sean X., and Xiuli Chao. 2010. Newsvendor bounds and heuristics for serial supply chains with regular and expedited shipping. *Naval Research Logistics* **57** (1) : 71-87.

Online Appendix: Shipping Consolidation with Delivery Deadline and Expedited Shipment Options

Proof of Proposition 1: For any optimal solution of $V_t(z)$, it is also feasible for $V_t(z+1)$. Thus, $V_t(z+1) \leq V_t(z)$. Also, extending the time-to-go horizon increases the total shipping cost, as the shipping cost is positive.

Proof of Theorem 2: Let $X_i \sim \text{Geometric}(\alpha)$ for all i. (We assume that $X_i's$ are i.i.d.) Consider a sufficiently long time horizon T. If we use the same threshold τ in all periods, then the whole selling horizon can be approximately decomposed into N random cycles $S_1, S_2, ..., S_N$, where $S_i = X_i + d - \tau$ and N is the smallest n such that $\sum_{i=1}^{n} S_i > T$. (Intuitively, $N-1$ is the number of shipments during T periods.) Note that N is a stopping time, so by Wald's equation, $\mathbf{E}\left[\sum_{i=1}^{N} S_i\right] = \mathbf{E}[N] \cdot \mathbf{E}[S_1] = \mathbf{E}[N]\left(\frac{1}{\alpha} + d - \tau\right)$. Since $\sum_{i=1}^{N-1} S_i \leq T < \sum_{i=1}^{N} S_i$, we have $(\mathbf{E}[N]-1)\left(\frac{1}{\alpha} + d - \tau\right) \leq T < \mathbf{E}[N]\left(\frac{1}{\alpha} + d - \tau\right)$. If T is sufficiently large, $\mathbf{E}[N]$ is approximately $T\left(\frac{1}{\alpha} + d - \tau\right)^{-1}$ and the average shipping costs is approximately $C(\tau, \alpha, d) := F(\tau)\left(\frac{1}{\alpha} + d - \tau\right)^{-1}$.

Proof of Proposition 2: Denote $F(x) = a - bx$ $(F(d) \geq 0)$, the long run cost is $C = \min_x C(x) = \min_x \frac{a-bx}{\frac{1}{\alpha}-1+d-x}$. We have $C'(x) = \frac{-b(\frac{1}{\alpha}-1+d-x)+(a-bx)}{(\frac{1}{\alpha}-1+d-x)^2} = \frac{b(1-\frac{1}{\alpha})+a-bd}{(\frac{1}{\alpha}-1+d-x)^2}$. (1) When $b(1-\frac{1}{\alpha})+a-bd \geq 0$ $(\alpha \geq \frac{b}{a-bd+b})$, $\tau^* = 1$. (2) When $b(1-\frac{1}{\alpha})+a-bd < 0$ $(\alpha < \frac{b}{a-bd+b})$, $\tau^* = d-1$.

Proof of Lemma 3: We have $C'_\tau = T\frac{F'(\tau)[\frac{1}{\alpha}+d-\tau]+F(\tau)}{(\frac{1}{\alpha}+d-\tau)^2}$. As $F(z)$ is first-order continuous, C'_τ is continuous. Next, we show that C'_τ is increasing in τ by showing $C'_\tau - C'_{\tau-\delta} \geq 0\ \forall \delta$. $C'_\tau - C'_{\tau-\delta} = T\frac{F'(\tau)[\frac{1}{\alpha}+d-\tau]+F(\tau)}{(\frac{1}{\alpha}+d-\tau)^2} - T\frac{F'(\tau-\delta)[\frac{1}{\alpha}+d-\tau+\delta]+F(\tau-\delta)}{(\frac{1}{\alpha}+d-\tau+\delta)^2} \geq T\frac{F'(\tau)[\frac{1}{\alpha}+d-\tau]+F(\tau)-F'(\tau-\delta)[\frac{1}{\alpha}+d-\tau+\delta]+F(\tau-\delta)}{(\frac{1}{\alpha}+d-\tau+\delta)^2}$. As the denominator is always positive, we evaluate the numerator. $F'(\tau)[\frac{1}{\alpha}+d-\tau]+F(\tau)-F'(\tau-\delta)[\frac{1}{\alpha}+d-\tau+\delta]+F(\tau-\delta) = [F'(\tau)-F'(\tau-\delta)](1/\alpha+d-\tau)+[F(\tau)-F(\tau-\delta)-\delta F'(\tau-\delta)] \geq 0$. The inequality follows as $F'(\tau)-F'(\tau-\delta) \geq 0$ and $F(\tau) \geq F(\tau-\delta)+\delta F'(\tau-\delta)$, since $F(x)$ is decreasing and convex. Next, we divide the analysis into two parts.

Scenario 1: If $F'(0)[1/\alpha+d]+F(0) < 0$: We discuss the changes of threshold in the following two cases: (a) $F'(d)1/\alpha+F(d) > 0$ and (b) $F'(d)1/\alpha+F(d) \leq 0$, for a specific α and $\tau = d$. **(a)** As C'_τ is continuous, $\exists \tau^* < d$, such that $F'(\tau^*)[1/\alpha+d-\tau^*]+F(\tau^*) = 0$ and τ^* is the smallest τ that satisfies this equation. For $\alpha' \geq \alpha$, $F'(\tau^*)[1/\alpha'+d-\tau^*]+F(\tau^*) > 0$. Thus, there exist $\tau' < \tau^*$, such that $F'(\tau')[\frac{1}{\alpha}+d-\tau']+F(\tau') = 0$. Therefore, when α increases, τ decreases. **(b)** $F'(d)1/\alpha+F(d) \leq 0$. As C'_τ is continuous and increasing, together with $F'(0)[1/\alpha+d]+F(0) < 0$, $F'(\tau)[1/\alpha+d-\tau]+F(\tau)$ must be negative for all $\tau \in [0,d]$. Thus $\tau^* = d$. Also, $F'(d)[1/\alpha]+F(d)$ increases in α, since $F'(d)$ is negative. For $\alpha' > \alpha$, if $F'(d)[1/\alpha']+F(d) > 0$,

then it is the same logic as in case (a) and the optimal τ^* moves to τ' where $\tau' < \tau^*$. If $F'(d)[1/\alpha'] + F(d) \leq 0$, then $\tau^* = d$. In either case, τ is non-increasing when α increases. **Scenario 2**: If $F'(0)[1/\alpha + d] + F(0) \geq 0$: As C'_τ is increasing with τ, $F'(\tau)[1/\alpha + d - \tau] + F(\tau) \geq 0$ for all $\tau \in [0, d]$. Thus, $\tau^* = 1$. For $\alpha' > \alpha$, it is easy to see that $F'(\tau)[1/\alpha + d - \tau] + F(\tau) \geq 0$ for all $\tau \in [0, d]$ also hold, as $F'(.)$ is negative. Thus, τ^* stays as 1. Thus, in either scenario, when α increases, τ decreases.

Proof of Proposition 3: For the first part, we argue the case $z'_A \leq z_A$ in detail. The other cases are similar. For any $V_t(z_A - 1, z_B, z_C)$, the same optimal shipping policy can be applied for $V_t(z_A, z_B, z_C)$. Thus, the optimal solution is a feasible one for $V_t(z_A, z_B, z_C)$ and $V_t(z_A - 1, z_B, z_C) \geq V_t(z_A, z_B, z_C)$. For the second part, a longer time horizon increases the total shipping cost, as the shipping cost is positive.

Proof of Lemma 5: We only focus on the first part, as the second one can be argued in a similar way. Suppose that $z_A, z_B < \infty$ and it is optimal to ship product B from W1. We will argue that it is also optimal to ship order type A together with B. We divide the analysis into two cases: (1) If $z_A \leq z_B$, consider a modified policy that does not ship order type B in the current period, but instead ships it at time $t' < t$ when type A is shipped. Clearly, we save the current shipment cost without adding new cost. (2) If $z_A > z_B$, then shipping type A together with type B in the current period does not increase shipping cost.

Proof of Lemma 7 We first provide the following two results, Lemma E1 and Lemma E2, which will be useful in the proof of Lemma 7.

LEMMA E1. *In symmetric case where* $F_1(x) = F_2(x) = F(x)\ \forall x$ *and* $\alpha_A = \alpha_C$, $V_t(z_A, \infty, \infty) - V_t(\infty, \infty, \infty) \geq F(z_A) - F(d)$, $V_t(\infty, z_B, \infty) - V_t(\infty, \infty, \infty) \geq F(z_B) - F(d)$, $V_t(\infty, \infty, z_C) - V_t(\infty, \infty, \infty) \geq F(z_C) - F(d)$. $V_t(\infty, z_B, z_C) - V_t(\infty, \infty, z_C) \geq F(z_B) - F(z_C)$, *for* $z_B \leq z_C$. $V_t(z_A, z_B, \infty) - V_t(z_A, \infty, \infty) \geq F(z_B) - F(z_A)$ *for* $z_B \leq z_A$.

Proof of Lemma E1: We prove the claims above by induction. For $t = 1$, $V_t(z_A, \infty, \infty) - V_t(\infty, \infty, \infty) = F(z_A) \geq F(z_A) - F(d)$ and it is easy to see that other inequalities also hold. Suppose the inequalities hold for all $t \leq t_0$. Then for $t = t_0 + 1$, we show the proof for the first inequality in detail. We know that $V_t(z_A, \infty, \infty) = \min C_{(x_A, x_B, x_C)}(z_A, \infty, \infty) = \min\{f(z_A x_A, \infty, \infty) + \mathbf{E}[V_{t-1}(\tilde{z}_A, \tilde{z}_B, \tilde{z}_C)]\}$ and $V_t(\infty, \infty, \infty) = \min C_{(x_A, x_B, x_C)}(\infty, \infty, \infty) = \min\{f(\infty, \infty, \infty) + \mathbf{E}[V_{t-1}(\tilde{z}_A, \tilde{z}_B, \tilde{z}_C)]\}$. For each (x_A, x_B, x_C), $C_{(x_A, x_B, x_C)}(z_A, \infty, \infty) - C_{(x_A, x_B, x_C)}(\infty, \infty, \infty) \leq F(z_A) - F(d)$, where the last inequality follows from induction hypothesis and simple algebra. Other inequalities can be shown in a similar way. □

Before stating Lemma E2, note that the DP formulation can be equivalently written as a pseudo-DP formulation as follows. Note that the shipping alternative of shipping only B from W1 or W2 is omitted by Lemma 5. For $t > 1$ and $z_A, z_B, z_C \geq 1$, we have: $V_t(z_A, z_B, z_C) = \min\{F_1(z_A) + \tilde{V}_t^1(\infty, z_B, z_C), F_1(\min\{z_A, z_B\}) + \tilde{V}_t^1(\infty, \infty, z_C), \tilde{V}_t^1(z_A - 1, z_B - 1, z_C - 1)\}$, where the corresponding alternatives are "Ship A from W1," "Ship A and B from W1," and "Do not ship from W1," respectively. $\tilde{V}_t^1(z_A, z_B, z_C) = \min\{F_2(z_C) + \tilde{V}_t^2(z_A, z_B, \infty), F_2(\min\{z_B, z_C\}) + \tilde{V}_t^2(z_A, \infty, \infty), \tilde{V}_t^2(z_A, z_B, z_C)\}$, where the corresponding alternatives are "Ship C from W2," "Ship C and B from W2," and "Do not ship from W2," respectively. $\tilde{V}_t^2(z_A, z_B, z_C) = \mathbf{E}[V_{t-1}(g_A(z_A), g_B(z_B), g_C(z_C))]$, where $g_X(z_X)$ is a random variable which equals $z_X - 1$ with probability $1 - \alpha_X$ and d with probability α_X. For $t = 1$ and $z_A, z_B, z_C \geq 1$, $V_1(z_A, z_B, z_C) = \min\{F_1(\min\{z_A, z_B\}) + F_2(z_C), F_1(z_A) + F_2(\min\{z_B, z_C\})\}$.

LEMMA E2. *In a symmetric case, where $F_1(x) = F_2(x) = F(x)\ \forall x$ and $\alpha_A = \alpha_C$: (1) If $z_A \geq \max\{z_B, z_C\}$, then $V_t(z_A - 1, z_B, z_C) - V_t(z_A, z_B, z_C) \geq F_1(z_A - 1) - F_1(z_A)$. (2) If $z_C \geq \max\{z_A, z_B\}$, then $V_t(z_A, z_B, z_C - 1) - V_t(z_A, z_B, z_C) \geq F_2(z_C - 1) - F_2(z_C)$.*

Proof of Lemma E2: This is an extension of Lemma 6. We prove it by induction. For $t = 1$, $V_t(z_A - 1, z_B, z_C) = F(z_A - 1) + F(z_C)$ and $V_t(z_A, z_B, z_C) = F(z_A) + F(z_C)$, where the inequality holds. Suppose it holds for $t \leq t_0$. Then, for $t = t_0 + 1$: (1) for $z_A \geq z_B \geq z_C$, "ship A and B from W1" is dominated by "ship B and C from W2." As in symmetric case $V_t(\infty, \infty, z_C) - V_t(z_A, \infty, \infty) = V_t(z_C, \infty, \infty) - V_t(z_A, \infty, \infty) \geq F(z_C) - F(z_A) \geq F(z_C) - F(z_B)$, $F(\min\{z_A, z_B\}) + V_t(\infty, \infty, z_C) \geq F(\min\{z_B, z_C\}) + V_t(z_A, \infty, \infty)$. Second, note that "ship C from W2" cannot be optimal, as $F(z_C) + V_t(z_A - 1, z_B, \infty) \geq F(z_C) + V_t(z_A - 1, \infty, \infty) = F(\min\{z_B, z_C\}) + V_t(z_A - 1, \infty, \infty)$. (2) for $z_A \geq z_C \geq z_B$, "Shipping C from W2" is dominated by "shipping both B and C from W2" as $F(z_B) + V_t(z_A, \infty, \infty) \leq F(z_C) + V_t(z_A, z_B, \infty)$, which follows from $V_t(z_A, z_B, \infty) - V_t(z_A, \infty, \infty) \geq F(z_B) - F(z_A) \geq F(z_B) - F(z_C)$. Also, "ship A and B from W1" is dominated by "ship B and C from W2" as $F(z_B) + V_t(\infty, \infty, z_C) \geq F(z_B) + V_t(z_A, \infty, \infty)$, which follows from $V_t(\infty, \infty, z_C) = V_t(z_C, \infty, \infty) \geq V_t(z_A, \infty, \infty)$. Thus, only three shipping alternatives need to be considered in the pseudo-DP: "Ship A from W1," "Ship C and B from W2" and "Do not ship." $V_t(z_A, z_B, z_C) = \min\{F(z_A) + V_t(\infty, z_B, z_C), F(\min\{z_B, z_C\}) + V_t(z_A, \infty, \infty), V_{t-1}(z_A - 1, z_B - 1, z_C - 1)\}$. Then, from induction hypothesis, Lemmas 6 and 11, $V_t(z_A - 1, z_B, z_C) - V_t(z_A, z_B, z_C) \geq F(z_A - 1) - F(z_A)$. This completes the proof. The $z_C \geq \max\{z_A, z_B\}$ part can be proved by similar logic. $\square$

Next, we show the six-boundary result by splitting the proof into three parts. In the first part, we show the existence of thresholds $\tau_{A,t}^{AB}(z_B, z_C)$ and $\tau_{C,t}^{BC}(z_A, z_B)$. The second part is for boundaries $\tau_{A,t}^{A}(z_B, z_C)$ and $\tau_{C,t}^{C}$, while the third part describes $\tau_{B,t}^{1}(z_A, z_C)$ and $\tau_{B,t}^{2}(z_A, z_C)$.

Part 1: The existence of thresholds $\tau_{A,t}^{AB}(z_B, z_C)$ is proved below in detail. Similar argument holds for $\tau_{C,t}^{BC}(z_A, z_B)$. We divide the proof into two cases, $z_A \le z_B$ and $z_A > z_B$. Case 1: $z_A \le z_B$. We show that, if for some (z_A, z_B, z_C), the optimal policy (x_A^*, x_B^*, x_C^*) is to ship A and B from W1 ($x_A^*, x_B^* = 1$), then for $(z_A - 1, z_B, z_C)$, the optimal policy is also to ship A and B from W1. For $(z_A - 1, z_B, z_C)$, $C_{(x_A^*, x_B^*, x_C^*)}(z_A - 1, z_B, z_C) = f(z_A - 1, z_B, z_C x_C^*) + \mathbf{E}[V_{t-1}(\tilde{z'}_A, \tilde{z}_B, \tilde{z}_C)]$, where $z'_A = z_A - 1$. As $x_A^* = 1$, $\mathbf{E}[V_{t-1}(\tilde{z'}_A, \tilde{z}_B, \tilde{z}_C)] = \mathbf{E}[V_{t-1}(\tilde{z}_A, \tilde{z}_B, \tilde{z}_C)]$. Thus, $C_{(x_A^*, x_B^*, x_C^*)}(z_A - 1, z_B, z_C) - C_{(x_A^*, x_B^*, x_C^*)}(z_A, z_B, z_C) = f(z_A - 1, z_B, z_C x_C^*) - f(z_A, z_B, z_C x_C^*) = F(z_A - 1) - F(z_A)$. Using (x_A^*, x_B^*, x_C^*) for $(z_A - 1, z_B, z_C)$ incurs $F(z_A - 1) - F(z_A)$ more cost than (z_A, z_B, z_C). From Lemma 6, we know that $V_t(z_A - 1, z_B, z_C) - V_t(z_A, z_B, z_C) \ge F(z_A - 1) - F(z_A)$, which indicates that using shipping policies other than (x_A^*, x_B^*, x_C^*) for $(z_A - 1, z_B, z_C)$ will incur at least additional $F(z_A - 1) - F(z_A)$ in cost compared to (z_A, z_B, z_C). Thus, (x_A^*, x_B^*, x_C^*) is the optimal policy of $(z_A - 1, z_B, z_C)$, which ships A and B from W1. Case 2: $z_A > z_B$. As (x_A^*, x_B^*, x_C^*) is optimal for (z_A, z_B, z_C), we know that $f(z_A, z_B, z_C x_C^*) = F(z_B) + F(z_C x_C^*)$ and $F(z_B) + F(z_C x_C^*) \le F(z_A) + F(\min\{z_B, z_C x_C^*\})$. For $(z_A - 1, z_B, z_C)$, $C_{(x_A^*, x_B^*, x_C^*)}(z_A - 1, z_B, z_C) = f(z_A - 1, z_B, z_C x_C^*) + \mathbf{E}[V_{t-1}(\tilde{z'}_A, \tilde{z}_B, \tilde{z}_C)]$ where $f(z_A - 1, z_B, z_C x_C^*) = F(z_B) + F(z_C x_C^*)$ as $F(z_B) + F(z_C x_C^*) \le F(z_A) + F(\min\{z_B, z_C x_C^*\}) \le F(z_A - 1) + F(\min\{z_B, z_C x_C^*\})$. Thus, $C_{(x_A^*, x_B^*, x_C^*)}(z_A - 1, z_B, z_C) = C_{(x_A^*, x_B^*, x_C^*)}(z_A, z_B, z_C)$. Following similar logic as in case 1, we know that using shipping policies other than (x_A^*, x_B^*, x_C^*) for $(z_A - 1, z_B, z_C)$ incurs higher cost than (z_A, z_B, z_C). Thus, (x_A^*, x_B^*, x_C^*) is the optimal policy of $(z_A - 1, z_B, z_C)$, which ships A and B from W1.

Part 2: The existence of thresholds $\tau_{A,t}^{1}(z_B, z_C)$ is proved in detail. Similar arguments hold for $\tau_{C,t}^{1}(z_A, z_B)$. We want to show that if, for some $V_t(z_A, z_B, z_C)$, the optimal policy (x_A^*, x_B^*, x_C^*) is to ship both A and B from W1 or to ship only A from W1 ($x_A^* = x_B^* = 1$ or $x_A^* = 1$), then for $V_t(z_A - 1, z_B, z_C)$, the optimal policy is also to ship both A and B from W1 or to ship only A from W1. For $x_A^* = x_B^* = 1$, it is already discussed in part 1. We next prove $x_A^* = 1$ part: We want to show that if, for some $V_t(z_A, z_B, z_C)$, the optimal policy (x_A^*, x_B^*, x_C^*) is to ship only A from W1 ($x_A^* = 1$), then for $V_t(z_A - 1, z_B, z_C)$, the optimal policy is also to ship both A and B from W1 or to ship only A from W1. Only $z_A > \max\{z_B, z_C\}$ need to be analyzed in this part, because in other two cases (1) and (2), shipping A from W1 incurs larger cost than shipping A and B from W1: (1) with $z_A \le z_B$, $F(z_A) + V_t(\infty, z_B, z_C) \ge F(\min\{z_A, z_B\}) + V_t(\infty, \infty, z_C)$. (2) with $z_C \ge z_A > z_B$,

$F(z_A)+V_t(\infty,z_B,z_C)\geq F(z_B)+V_t(\infty,\infty,z_C)$, as $V_t(\infty,z_B,z_C)-V_t(\infty,\infty,z_C)\geq F(z_B)-F(z_C)\geq F(z_B)-F(z_A)$ by Lemma E1. With $z_A>\max\{z_B,z_C\}$, as $x_A^*=1$, $\mathbf{E}[V_{t-1}(\tilde{z'}_A,\tilde{z}_B,\tilde{z}_C)]=\mathbf{E}[V_{t-1}(\tilde{z}_A,\tilde{z}_B,\tilde{z}_C)]$. Thus, $C_{(x_A^*,x_B^*,x_C^*)}(z_A-1,z_B,z_C)-C_{(x_A^*,x_B^*,x_C^*)}(z_A,z_B,z_C)=f(z_A-1,z_Bx_B^*,z_Cx_C^*)-f(z_A,z_Bx_B^*,z_Cx_C^*)=F(z_A-1)-F(z_A)$. From Lemma E2, we know that $V_t(z_A-1,z_B,z_C)-V_t(z_A,z_B,z_C)\geq F(z_A-1)-F(z_A)$, which indicates that using shipping policies other than (x_A^*,x_B^*,x_C^*) for (z_A-1,z_B,z_C) will incur cost at least $F(z_A-1)-F(z_A)$ higher than (z_A,z_B,z_C). Thus, (x_A^*,x_B^*,x_C^*) is the optimal policy of (z_A-1,z_B,z_C), which ships A and B from W1.

Part 3: The existence of thresholds $\tau_{B,t}^1(z_A,z_C)$ is proved in detail. Similar arguments hold for $\tau_{B,t}^2(z_A,z_C)$. We want to show that if, the optimal policy (x_A^*,x_B^*,x_C^*) is to ship both A and B from W1 ($x_A^*=x_B^*=1$), then for $V_t(z_A-1,z_B,z_C)$, the optimal policy is also to ship both A and B from W1. Only $z_A<z_C$ need to be considered in this part. Otherwise, with $z_A\geq z_C$, shipping both A and B from W1 is dominated by shipping both B and C from W2. This is because $F(\min\{z_A,z_B\})+V_t(\infty,\infty,z_C)\leq F(\min\{z_B,z_C\})+V_t(z_A,\infty,\infty)$, as $V_t(\infty,\infty,z_C)-V_t(z_A,\infty,\infty)=V_t(\infty,\infty,z_C)-V_t(\infty,\infty,z_A)\geq F(z_C)-F(z_A)\geq F(\min\{z_A,z_B\})-F(\min\{z_B,z_C\})$ (the first inequality is from Lemma E1). We divide the proof into two cases: (1) $z_B\leq z_A<z_C$. (2) $z_A<\min\{z_B,z_C\}$. **Case 1**: $z_B\leq z_A<z_C$. $C_{(x_A^*,x_B^*,x_C^*)}(z_A,z_B-1,z_C)-C_{(x_A^*,x_B^*,x_C^*)}(z_A,z_B,z_C)=f(z_A,z_B-1,z_Cx_C^*)-f(z_A,z_B-1,z_Cx_C^*)=F(z_B-1)-F(z_B)$, as in symmetric case $F(\min\{z_A,z_B-1\})+V_t(\infty,\infty,z_C)-F(\min\{z_A,z_B\})+V_t(\infty,\infty,z_C)=F(z_B-1)-F(z_B)$. From Lemma 6, we know that $V_t(z_A,z_B-1,z_C)-V_t(z_A,z_B,z_C)\geq F(z_B-1)-F(z_B)$, which indicates that using shipping policies other than (x_A^*,x_B^*,x_C^*) for (z_A,z_B-1,z_C) will incur at least additional cost of $F(z_B-1)-F(z_B)$ compared to (z_A,z_B,z_C). Thus, (x_A^*,x_B^*,x_C^*) is the optimal policy of (z_A,z_B-1,z_C), which ships A and B from W1. **Case 2**: $z_A<\min\{z_B,z_C\}$. $C_{(x_A^*,x_B^*,x_C^*)}(z_A,z_B-1,z_C)-C_{(x_A^*,x_B^*,x_C^*)}(z_A,z_B,z_C)=f(z_A,z_B-1,z_Cx_C^*)-f(z_A,z_B-1,z_Cx_C^*)=0$. From Proposition 3, we know that $V_t(z_A,z_B-1,z_C)\geq V_t(z_A,z_B,z_C)$, which indicates that for (z_A,z_B-1,z_C) using shipping policies other than (x_A^*,x_B^*,x_C^*) will incur higher cost than (z_A,z_B,z_C). Thus, (x_A^*,x_B^*,x_C^*) is the optimal policy of (z_A,z_B-1,z_C), which ships A and B from W1.

Proof of Lemma 8 First, we show a result of the effect of α on $V_t(z_A,z_B,z_C)-V_{t-1}(z_A,z_B,z_C)$, which will be useful in the proof of Lemma 8. Denote the cost-to-go function as $\bar{V}_t(z_A,z_B,z_C)$ and $V_t(z_A,z_B,z_C)$ for $\bar{\alpha}$ and α, respectively. Note that under optimal stationary policy, $V_t(z_A,z_B,z_C)-V_{t-1}(z_A,z_B,z_C)$ converges to the expected one-period cost. For $\bar{\alpha}\geq\alpha$, the expected one-period cost of $\bar{\alpha}$ is larger than that of α. This is because on average, more orders arrive per period in the case of $\bar{\alpha}$, which incurs more shipping cost. Thus,

$\bar{V}_t(z_A, z_B, z_C) - \bar{V}_{t-1}(z_A, z_B, z_C) \geq V_t(z_A, z_B, z_C) - V_{t-1}(z_A, z_B, z_C)$. Next, we show the proof of $\tau_A^A(z_B, z_C)$ in detail. Other cases are similar. For any z_B, z_C and $z_A = \tau_A^A\{z_B, z_C\}$, $V_t(z_A, z_B, z_C) = F_1(z_A) + V_t(\infty, z_B, z_C) \leq V_{t-1}(z_A - 1, z_B - 1, z_C - 1)$, where we can write $V_t(z_A, z_B, z_C)$ recursively (as a function of $V_t(\infty, z_B, z_C)$) because under optimal policy it is optimal to not ship in $V_t(\infty, z_B, z_C)$. Then consider $\bar{V}_t(z_A, z_B, z_C)$. Following $\bar{V}_t(z_A, z_B, z_C) - \bar{V}_{t-1}(z_A, z_B, z_C) \geq V_t(z_A, z_B, z_C) - V_{t-1}(z_A, z_B, z_C)$ and $\bar{V}_1(z_A, z_B, z_C) = V_1(z_A, z_B, z_C)$ $(\forall z_A, z_B, z_C)$, we have $\bar{V}_t(\infty, z_B, z_C) - V_t(\infty, z_B, z_C) \geq \bar{V}_{t-1}(z_A - 1, z_B - 1, z_C - 1) - V_{t-1}(z_A - 1, z_B - 1, z_C - 1)$. Thus, shipping only A (the policy corresponding to $\tau_A^A\{z_B, z_C\}$) is not necessarily the policy with smallest cost, which indicates the decreasing of $\tau_A(z_B, z_C)$.

Proof of Theorem 4: Before we show the proof, we first introduce notations for the changes of states across periods. For any state i in period t, we denote its predecessor in period $t+1$ as $p(i)$ and successors in period $t-1$ as $s(i)$. We start from proving parts 1 and 2. The states for $d = 3$ are (∞). (1), (2), (3), (1, 2), (1, 3), (2, 3) and (1, 2, 3). We prove, by induction, that in (1, 2), (1, 3), (2, 3) and (1, 2, 3), it is not optimal to ship only partial of the orders. For $t = 1$, by definition, it is optimal to ship all of the pending orders. Suppose it is not optimal to ship only partial of the orders for $t \leq t' - 1$ periods. In other words, it is optimal to either ship all orders or do not ship in states (1, 3), (1, 2), (1, 2, 3) and (2, 3). We only list these four cases, as in other cases it is natural to either ship the order or not. For $t = t'$, we consider the following 4 cases. **Case 1: State** $(1, 3)$**.** Suppose in period t under the optimal policy π^* it is optimal to ship only order 1 in state $(1, 3)$. Note that $p(1, 3) = (2)$ in period $t+1$ and the optimal policy for (2) has to be "Do not ship." Next, we argue that "Do not Ship" cannot be an optimal policy for (2). If (2) is not shipped in period $t+1$, the state in period t becomes (1) with probability α and $(1, 3)$ with probability $1 - \alpha$. In either case, order 1 is shipped, resulting in a higher cost than in the previous period. Now, if the optimal policy of $(1, 3)$ is to ship both orders but in separate packages, the above argument also holds. Thus, if state $(1, 3)$ is reachable, the optimal policy is to ship both orders in one package. $V_t(1, 3) = C(1, 3) + V_t(\infty) \leq C(1) + V_t(3)$, where $C(1, 3) = F(1) + 2v(1) \leq F(1) + v(1) + F(3) + v(3)$, and the following inequality holds, $v(1) \leq F(3) + v(3)$. Note that the equation $(v(1) \leq F(3) + v(3))$ does not impose any assumption about the relation between $F(.)$ and $v(.)$. It only states that, if state $(1, 3)$ is reachable, then this equation must hold. In other words, if this equation does not hold, then state $(1, 3)$ cannot be reached under the optimal policy. **Case 2: State** $(1, 2)$**.** Suppose that under the optimal policy π^*, it is optimal to ship only order 1 in state $(1, 2)$ in period t. Then, in period $t - 1$, the remaining order (2) has successor $s(2) = \{(1, 3), (1)\}$ with probability α and

$1-\alpha$, respectively. Thus, state (1,3) is reachable and inequality $(v(1) \leq F(3)+v(3))$ holds. Then, $C(1,2) = \min\{F(1)+2v(1), F(1)+v(1)+F(2)+v(2)\} = F(1)+2v(1) = C(1,3)$, as $F(1)+2v(1) \leq F(1)+v(1)+F(3)+v(3) \leq F(1)+v(1)+F(2)+v(2)$, where the first inequality follows from $v(1) \leq F(3)+v(3)$. Also, note that $V_t(1,2) = \min\{C(1,2)+V_t(\infty), C(1)+V_t(2)\} = C(1,2)+V_t(\infty)$, as $C(1,2)+V_t(\infty) = C(1,3)+V_t(\infty) \leq C(1)+V_t(3) \leq C(1)+V_t(2)$. Thus, shipping both order 1 and 3 in one package incurs smaller cost than π^*, which contradicts the optimality of π^*. Thus, if state $(1,2)$ is reachable, the optimal policy is to ship both orders in one package. $V_t(1,2) = C(1,2)+V_t(\infty)$, where $C(1,2) = F(1)+2v(1)$. **Case 3: State** $(1,2,3)$**.** There are 4 shipment alternatives for state $(1,2,3)$: ship 1 alone, ship 1 and 2, ship 1 and 3 and ship all orders. We argue that the first three alternatives cannot be optimal. First, suppose that under the optimal policy π^* it is optimal to ship only order $(1,2)$ in state $(1,2,3)$ in period t. Note $p(1,2,3) = (2,3)$ where the optimal policy for $(2,3)$ should be "Do not ship." Thus, $s(2,3) = \{(1,2,3),(1,2)\}$ with probability α and $1-\alpha$, respectively. As $(1,2)$ is reachable, from the results in Case 2, $C(1,2) = F(1)+2v(1)$. Note that in both cases of $(1,2,3)$ and $(1,2)$, orders $(1,2)$ are shipped in one package and incur cost $F(1)+2v(1)$. Consider policy $\tilde{\pi}$ which chooses to ship in state $(2,3)$ in period $t+1$ and keeps other decisions the same as π^*. The cost of $\tilde{\pi}$ is $F(2)+2v(2)$, which is smaller than that of π^*, which contradicts the optimality of π^*. Also, note that, if the optimal policy for $(1,2,3)$ is to ship order 1 and 2 in package one and order 3 in package two, the argument also holds. Second, suppose it is optimal to ship 1 and 3 in state $(1,2,3)$ in period t. If (2) is not shipped in period t, the state in period $t-1$ becomes $(1,3)$ with probability α and (1) with probability $1-\alpha$. As $(1,3)$ is reachable, $v(1) \leq F(3)+v(3)$ holds. Then, $C(1,3) = F(1)+2v(1) \leq F(1)+v(1)+F(3)+v(3)$. Thus, 1 and 3 should be shipped in the same package. Consider policy $\tilde{\pi}$ that ships orders 1 and 2 in period t and ships order 3 following the same policy as order 2 in policy π^*. $\tilde{\pi}$ incurs the same cost in period t but lower cost in future period. It contradicts the optimality of policy π^*. Third, suppose under the optimal policy π^* it is optimal to ship only order (1) in state $(1,2,3)$ in period t. Thus, in period $t-1$, $s(2,3) = \{(1,2,3),(1,2)\}$ with probability α and $1-\alpha$, respectively. From induction hypothesis, it is optimal to ship all orders in state $(1,2,3)$ and $(1,2)$ in period $t-1$. Thus, the remaining order 2 in period t incurs at least variable cost $v(1)$ in period $t-1$. Consider policy $\tilde{\pi}$ which chooses to ship $(1,2)$ in state $(1,2,3)$ in period t, and keeps other decisions the same as π^*. $\tilde{\pi}$ incurs additional cost of $v(1)$ in period t, while the cost decreases at least by $v(1)$ in period $t-1$. It contradicts the optimality of policy π^*. Thus, if state $(1,2,3)$ is reachable, the optimal policy is to ship all orders. Whether to ship them in one package or in separate packages depends on the relationship of $F(.)$

and $v(.)$: $C(1,2,3) = \min\{F(1)+3v(1), F(1)+v(1)+F(2)+2v(2), F(1)+v(1)+F(2)+v(2)+F(3)+v(3)$.

Case 4: State $(2,3)$**.** We show that shipping either (2) or (3) along can not be optimal. First, suppose under the optimal policy π^* it is optimal to ship only order 2 in state $(2,3)$ in period t. In period $t-1$, the remaining order (3) has successor $s(3) = \{(2,3),(2)\}$ with probability α and $(1-\alpha)$, respectively. From induction hypothesis, in state $(2,3)$, it is optimal to either ship both orders or not to ship. Thus, the remaining order (3) in period t will be shipped with other orders arriving in later periods, which incurs at least variable cost $v(2)$. Consider policy $\tilde{\pi}$ which ships $(2,3)$ in period t and keep other decisions the same as in policy π^*. $\tilde{\pi}$ incurs $v(2)$ higher cost in period t while decrease at least $v(2)$ cost in later periods. It contradicts the optimality of policy π^*. Second, suppose under the optimal policy π^* it is optimal to ship only order 3 in state $(2,3)$ in period t. In period $t-1$, the remaining order (2) has successor $s(2) = \{(1,3),(1)\}$ with probability α and $(1-\alpha)$, respectively. It is easy to see that exchange the policy of order 2 and 3 in period t incurs $v(2)-v(3)$ higher cost in period t while decreases at least $v(2)-v(1)$ cost in period $t-1$. It contradicts with the optimality of π^*. Thus, if state $(2,3)$ is reachable, the optimal policy is to ship both orders. Whether to ship them in one package or in separate packages depends on the relation between $F(.)$ and $v(.)$: $C(2,3) = \min\{F(2)+2v(2), F(2)+v(2)+F(3)+v(3)\}$.

Finally, we proof the third part of Theorem 4 using the result from the first and second parts of Theorem 4. Note that it is sufficient to consider the following three scenarios: Scenario 1, if it is optimal to ship for state (1), then it is optimal to ship in states $(1,2)$ and $(1,3)$; Scenario 2, if it is optimal to ship for state $(1,2)$ or $(1,3)$, then it is optimal to ship in state $(1,2,3)$; Scenario 3, if it is optimal to ship for state (2), then it is optimal to ship in state $(2,3)$. The first two scenarios are obvious, as the order needs to be shipped for $z_1 = 1$. Thus, we only need to prove the third scenario. As the optimal policy for $V_t(2)$ $(\forall t > 1)$ is to ship, $F(2)+v(2)+V_t(\infty) \leq \alpha V_{t-1}(1) + (1-\alpha)V_{t-1}(1,3)$. Then, for state $(2,3)$, $F(2)+2v(2)+V_t(\infty) \leq V_{t-1}(1)+(1-\alpha)V_{t-1}(1,3)+v(2) \leq \alpha V_{t-1}(1,2,3)+(1-\alpha)V_{t-1}(1,2)$. Thus, it is also optimal to ship in state $(2,3)$.

Proof of Theorem 5: Denote the packages shipped under any policy π in T periods as p_i^π $(i \leq k^\pi$, where k^π is the total number of packages). In package p_i^π, denote the smallest slack time as $z(p_i^\pi)$ and the number of orders in the package as $m(p_i^\pi)$. The total cost of any policy π is $E[\sum_{i \leq k^\pi} F(z(p_i^\pi)) + m(p_i^\pi)\ v(z(p_i^\pi))]$. **For the first bound**: We derive the relation between the costs C^* of the optimal policy π^* and the costs C_F of the policy π_F which considers only fixed cost. $C^* = E[\sum_{i \leq k^{\pi^*}} F(z(p_i^{\pi^*})) + m(p_i^{\pi^*})\ v(z(p_i^{\pi^*}))] \geq$

$E[\sum_{i\le k^{\pi^*}} F(z(p_i^{\pi^*})) + \gamma F(z(p_i^{\pi^*}))] = E[(1+\gamma) \sum_{i\le k^{\pi^*}} F(z(p_i^{\pi^*}))] \ge (1+\gamma)E[\sum_{i\le k^{\pi_F}} F(z(p_i^{\pi_F}))]$. The first inequality follows from $v=\gamma F$ and the fact that there must be at least one order in each packages. The second inequality follows from the optimality of π_F, which considers only fixed cost. Thus, $C^* \ge (1+\gamma)C_F$. Let $C(\pi_F)$ denote the cost of applying π_F in the case with both fixed and variable cost. Then, $C(\pi_F) \ge C^* \ge (1+\gamma)C_F$. As π_F is one-threshold(τ) policy, the cost of $C(\pi_F)$ and C_F can be derived explicitly as $C_F = \frac{T}{\frac{1}{\alpha}+d-\tau}F(\tau)$ and $C(\pi_F) = \frac{T}{\frac{1}{\alpha}+d-\tau}[F(\tau)+\mathbb{E}(m)v(\tau)]$ where $\mathbb{E}(m) = 1+(d-\tau)\alpha$. Plugging in the inequality $C(\pi_F) \ge C^* \ge (1+\gamma)C_F$ and by simple algebra, $\frac{C(\pi_F)-C^*}{C^*} \le \frac{1}{1+\gamma}\frac{C(\pi_F)}{C_F} - 1 \le \frac{\gamma(d-\tau)\alpha}{1+\gamma}$. **For the second bound**: We derive the relation between C^* and the costs C_v of the policy π_v which considers only variable cost. $C^* = E[\sum_{i\le k^{\pi^*}} F(z(p_i^{\pi^*})) + m(p_i^{\pi^*})v(z(p_i^{\pi^*}))] = E[\sum_{i\le k^{\pi^*}} \frac{1}{\gamma}v(z(p_i^{\pi^*})) + m(p_i^{\pi^*})v(z(p_i^{\pi^*}))] \ge E[\sum_{i\le k^{\pi^*}} \frac{1}{\gamma}v(d) + m(p_i^{\pi^*})v(d)] \ge \frac{v(d)}{\gamma}\frac{T}{\frac{1}{\alpha}+d} + C_v$. The last inequality follows from the fact that the expected number of packages shipped in the long run is $\frac{T}{\frac{1}{\alpha}+d}$ and the optimality of C_v, which considers only variable cost. Denote the cost of applying π_v to the case with both fixed and variable cost as $C(\pi_v)$. Then, $C(\pi_v) \ge C^* \ge C_v + \frac{v(d)}{\gamma}\frac{T}{\frac{1}{\alpha}+d}$. Note that $C(\pi_v)$ and C_v can be written explicitly as orders are shipped upon arrivals: $C(\pi_v) = \frac{T}{\frac{1}{\alpha}}[F(d)+v(d)] = T\alpha[F(d)+v(d)]$ and $C_v = T\alpha v(d)$. Plugging $C(\pi_v)$ and C_v into the above inequality and by simple algebra, $\frac{C(\pi_v)-C^*}{C^*} \le \frac{C(\pi_v)-C_v-\frac{v(d)}{\gamma}\frac{T}{\frac{1}{\alpha}+d}}{C_v-\frac{v(d)}{\gamma}\frac{T}{\frac{1}{\alpha}+d}} = \frac{k}{\gamma+1-k}$, where $k = \frac{d\alpha}{1+d\alpha}$.