Lead-Time-Constrained Middle-Mile Consolidation Network Design with Fixed Origins and Destinations

Lacy M. Greening*, Mathieu Dahan, Alan L. Erera
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia
{lacy.greening@gatech.edu, mathieu.dahan@isye.gatech.edu, alan.erera@isye.gatech.edu}

Many large e-commerce retailers move sufficient freight volumes to operate private middle-mile consolidation networks for order fulfillment, transporting customer shipments from stocking locations to last-mile delivery partners in consolidated loads to reduce freight costs. We study a middle-mile network design optimization problem with fixed origins and destinations to build load consolidation plans that minimize cost and satisfy customer shipment lead-time constraints. We propose models that extend traditional flat network service network design problems to capture waiting delays between load dispatches and ensure that shipment lead-time requirements are satisfied with a desired probability. We approximate these chance constraints using hyperparameterized linear constraints, resulting in new mixed-integer programs (MIPs) for service network design. To find high-quality solutions to the proposed MIPs, we develop an effective integer-programming-based local search (IPBLS) heuristic that iteratively improves a solution by optimizing over a smartly selected subset of commodities. For the largest problem instances, we propose a two-phase IPBLS heuristic that first utilizes a simplified, restricted MIP that constrains leg waiting delays individually. Computational experiments using data from a large U.S.-based e-commerce partner demonstrate the significant impact of tight lead-time constraints on the structure of the consolidation network designs and their concomitant operating costs. Notably, tighter constraints lead to solutions with increased shipment consolidation and higher dispatch frequencies on selected key transportation lanes. Such solutions trade off higher shipment transit times with significantly reduced shipment waiting times to meet lead-time constraints at lower cost.

Key words: e-commerce logistics; service network design; middle mile; local search;
History: This article was first submitted on September 18, 2022.

1. Introduction

E-commerce retailing models such as ship-to-home and ship-to-store require that retailers fulfill orders to customers on demand. E-retailers typically ship goods direct to customers from one or more *fulfillment centers* (FCs) and/or from product vendor locations when drop shipping. When shipping direct, e-retailers arrange shipments with third party carriers from origin stocking locations through to final customer delivery locations. However, large firms may be able to generate substantial cost savings by alternatively building *consolidated loads* with many shipments outbound

* Corresponding author.

from some stocking locations into other facilities and potentially transferring those shipments into subsequent consolidated loads prior to last-mile delivery. Such a system of consolidated loads is a private *middle-mile* network and the design of these networks is the focus of this paper.

Amazon, Wayfair, and The Home Depot are examples of U.S. shippers who actively manage large-scale middle-mile networks for e-commerce fulfillment. Amazon was one of the first to do so, and they currently build outbound truckloads of packages from FCs to both dedicated intermediate sort centers and direct to Amazon Prime delivery stations or third party parcel carrier facilities (Leonard, M. 2021). Wayfair has also recently made middle-mile investments and coordinates logistics services for many of their vendors for direct-to-customer shipments; in 2020, 90% of their U.S. large-parcel orders flowed through a private middle-mile network (Wayfair 2021). As a final example, The Home Depot is expanding its middle-mile fulfillment system as part of their One Supply Chain initiative (The Home Depot 2021), leveraging new and existing fulfillment center locations as intermediate consolidation locations. Each of these companies uses a private network to reduce outbound transportation cost while simultaneously speeding up order transit times and decreasing their reliance on third party logistics providers.

In this paper, we consider a specific middle-mile network planning problem faced by some large e-retailers. Here, the retailer must ship orders over time from known origin stocking locations (FCs or vendor locations) to known destination last-mile distribution (LMD) facilities. Examples of such LMD facilities may be those operated by package transportation companies or postal services (e.g., UPS), branded delivery subsidiaries (e.g., Amazon Prime), and/or less-than-truckload (LTL) carriers or local large-and-bulky delivery companies. Since customer orders each have a promised delivery time, each shipment should move from its origin to its LMD destination to meet a time deadline. To minimize costs, the retailer consolidates shipments when appropriate into larger loads (for example, truckloads or larger less-than-truckload shipments) prior to dispatch. These consolidated loads are then outsourced to third-party carriers for transportation. The planning problem then is to determine a joint set of shipment paths and load dispatches that move customer shipments from origins to destinations at minimum cost given delivery time requirements.

This middle-mile network planning problem is similar to service network design problems faced by consolidation trucking carriers, like LTL and package trucking firms. Traditional deterministic *flat network* service network design (SND) optimization models for trucking were first developed to configure such networks (see the review in Bakir *et al.*, 2021). In flat network models, shipment demand is modeled using average flow volumes, measured in total shipment size per time for each origin-destination pair, and capacity decision variables model the number of loads to be dispatched between facilities per time (i.e., dispatch *frequencies*). However, since these so-called flow and load planning models were developed originally for systems where origin-to-destination transit time standards were longer, they typically used simple lower bounds on weekly load dispatch frequencies to construct plans that would approximately meet these standards. Today, LTL and package carriers often promise tighter delivery times and in response, researchers have explored using detailed *time-expanded network models* for service planning that can model lead-time constraints more precisely. Since LTL and package carriers use sorting facilities that operate only a few dedicated sorting and dispatching periods each day, time-expanded networks with a large but manageable number of time periods provide a natural modeling approach for such operations (Erera *et al.* 2013). Nevertheless, such models tend to lead to extremely large integer programming optimization problems that are difficult to solve.

E-retailer middle-mile networks differ from carrier trucking networks, most notably in that shipments may be picked, packed, and ready to ship at origins continuously during operations. Subsequently, outbound loads in middle-mile networks may become ready for dispatch to various next destinations at many possible times each day. Accurate modeling of load dispatch times in middle-mile networks may require a larger set of time-space nodes and arcs, but time-expanded models may still be useful in middle-mile consolidation planning. In this paper, however, we explore a different modeling idea that reverts to flat network models that represent time as a continuum.

Specifically, this paper develops an approach that extends the use of traditional flat network models as the underlying infrastructure for a network design mixed-integer program (MIP) to create middle-mile planning models amenable to exact and heuristic solution approaches. Importantly, we show how to add probabilistic constraints on shipment lead times to such MIP models, where the lead time of a shipment may include transportation travel time and both fixed transfer processing time and potential additional waiting time between load dispatches. To summarize the primary contributions of this work, we:

- develop a new, effective MIP model, denoted the *middle-mile consolidation problem with waiting times* (MMCW), for consolidation network design that captures the time shipments spend waiting at transfer facilities within probabilistic lead-time constraints;
- approximate the chance constraints on shipment lead times using hyperparameterized nonlinear constraints and reformulate them as linear constraints using binary variables;
- build a powerful integer-programming-based (IP-based) local search heuristic with novel randomized search neighborhoods tailored to middle-mile consolidation problems and demonstrate its effectiveness for solving small- to medium-sized problem instances;
- develop a simpler, restricted MIP network design model, denoted the *middle-mile consolidation with allocated waiting delay* (MMCW-A), that allocates fractions of the total allowable waiting time for each possible shipment path in advance to individual path legs, dramatically reducing the size of the MIP formulation and producing high-quality starting solutions for the MMCW model when solving large instances;

- show that a two-phase IP-based local search heuristic that first searches a restricted solution space using the MMCW-A model before transitioning to using the MMCW model substantially outperforms an approach that relies solely on the MMCW model for the largest instances most similar to networks operated in practice by large U.S. retailers; and
- demonstrate the significant impact of more conservative lead-time constraints on the structure of the network designs produced for realistic instances and on the resulting middle-mile network operating costs.

The remainder of this paper is organized as follows. In Section 2, we discuss relevant literature. In Section 3, we formulate the lead-time-constrained middle-mile consolidation network design problem using two methods to model waiting delays. In Section 4, we develop single- and two-phase IP-based local search heuristics to solve the design problems. In Section 5, we present results from a computational study that highlight the impact of lead-time constraints on the resulting network designs and the effectiveness of our solution approaches. Finally in Section 6, we make concluding remarks and discuss potential areas of future work.

2. Literature Review

The consolidation network design problems faced by large e-retailers share many similarities with flow and load planning SND problems for consolidation trucking carriers, such as less-than-truckload (LTL) or package carriers (Bakir *et al.* 2021). We refer the reader to Crainic 2000 and Wieberneit 2008 for broad reviews of SND in transportation. Early trucking SND work focused on using flat (static) network models of a set of terminals with opportunities for consolidated truckloads to be dispatched between them represented by arcs with flows. Initial arc-based models specified minimum weekly truckload frequencies on arcs with positive truck flows to control waiting delays and ensure a minimum service level (Powell & Sheffi 1983, Powell 1986, Powell & Koskosidis 1992). These papers did not attempt to convert their formulations into linear MIP models and solve them; instead, local improvement heuristics are proposed that improve plans by sequentially adding and dropping facility-to-facility arcs to and from the network in an attempt to reduce costs. A different stream of early research proposed path-based flat network models for rail freight SND tactical planning applications (Crainic *et al.* 1984, Crainic & Rousseau 1986) and then adapted them for LTL transportation (Crainic & Roy 1988). These models select rail services to offer and their respective frequencies to meet demand requirements. To ensure that shipments are not delayed excessively, the proposed models included a nonlinear average waiting delay penalty in the objective function; because the resulting nonlinear MIPs are intractable, a decomposition-based algorithm is proposed to iteratively improve the plan by alternating between optimizing service frequencies with fixed flows and optimizing flows with fixed service frequencies. In our work herein, we will also propose a model that ensures a minimum service level by modeling waiting delays as a function of service frequencies. However, instead of setting simple lower bounds on arc frequencies or penalizing waiting delay in the objective, we develop chance constraints to ensure with a certain probability (or lower bound service guarantee) that shipments reach destinations within the promised lead time. We then solve the resulting MIP models exactly for small instances or to within a provable optimality gap for larger instances using a heuristic approach.

More recent work in flow and load planning focuses on more detailed modeling of the time shipments spend moving between origins and destinations by using time-expanded network models that explicitly capture when loads are to be dispatched, often referred to as scheduled service network design (SSND). Variants of SSND problems have been studied, including those that model empty resource management, stochastic shipment volumes and travel times, platooning, etc. (Lin 2001, Pedersen *et al.* 2009, Andersen *et al.* 2009, Lium *et al.* 2009, Bai *et al.* 2014, Zhu *et al.* 2014, Crainic *et al.* 2016, Demir *et al.* 2016, Scherr *et al.* 2019, Wang & Qi 2020). For larger networks and planning horizons, time-expanded models and the associated SSND MIPs become very large and difficult to solve. Models and heuristic solution methods of this type for planning trucking consolidation networks are introduced in Jarrah *et al.* (2009), Erera *et al.* (2013), and Lindsey *et al.* (2016). Each requires a heuristic approach to solve large-scale instances. Specifically, Jarrah *et al.* (2009) propose decomposition and slope scaling techniques and Erera *et al.* (2013) and Lindsey *et al.* (2016) propose approaches that use restricted and tailored integer programs, respectively, to find improving solutions. To produce plans of high quality, such models often rely on a fine discretization of time to accurately capture shipment consolidation opportunities; the quality of solutions may improve as the time windows narrow, but at the expense of computational challenges related to solving large and difficult MIPs. Other recent work has developed approaches that dynamically determine the exact times that dispatches should occur, and thus do not require specifying a time discretization in advance (Boland *et al.* 2017, Hewitt 2019, Boland *et al.* 2019, Scherr *et al.* 2020, Marshall *et al.* 2021, Hewitt 2022). These so-called dynamic discretization discovery approaches remain computationally expensive and have been shown to be effective primarily for networks with fewer than 50 nodes, 1,000 arcs, and at most 1,000 origin-destination pairs.

The modeling approach we develop in this paper for shipment waiting times in lead-time constraints is also similar to work found in the public transit literature. Using service headway (i.e., the inverse of frequency) to model passenger waiting times is common in work that addresses public transit systems (Mauttone *et al.* 2021). For example, when considering passengers arriving at a transit station according to a stationary stochastic process with independent increments, it is well known that the expected waiting time for each passenger until the arrival of a vehicle (i.e., bus) is equal to one-half of the vehicle dispatch headway when this headway is constant (Daganzo 1997). A problem related to public transit network design is the network assignment problem where models attempt to predict how passengers might jointly choose routing strategies across a network to move from origins to destinations, minimizing their traveling and waiting times. Spiess & Florian (1989) model average waiting time such problems as the inverse of total outbound departure frequency assuming exponential headways. Bouzaïene-Ayari *et al.* (2001) provide a review of this literature; more complex models of average waiting time have been proposed in other papers but the functions still generally include a term that is proportional to the inverse of departure frequencies. Cancela *et al.* (2015) extend passenger assignment models into transit design models that minimize passenger waiting times by assigning frequencies for selected services. To address the nonlinearity when modeling waiting delay as the inverse of frequency, they use a discrete set of frequency options for each service, of which one is assigned using a binary indicator variable; we adopt a similar modeling idea in this paper. Across this literature, we are not aware of other work that develops linearized chance constraints for total waiting time for multiple-leg trips.

Finally, although we develop tractable MIP models in this paper for lead-time-constrained middle-mile network design, we can obtain much better solutions to these models using less computation time by employing an effective heuristic solution approach known as IP-based local search (Hwang *et al.* 2011). IP-based local search is a MIP solution approach that combines exact and heuristic approaches (Franceschi *et al.* 2006, Savelsbergh & Song 2008, Hewitt *et al.* 2010). In this framework, a restricted version of the full MIP is solved at each iteration in an attempt to improve an incumbent solution (Erera *et al.* 2013, Lindsey *et al.* 2016). Specifically, at each iteration, a subset of variables are selected to be freed for optimization while the remaining are fixed to the current solution. The subset of variables to free is selected using defined search neighborhoods specific to the problem, such as freeing all variables associated with freight destined for a single terminal when optimizing an LTL load plan with an in-tree structure (Erera *et al.* 2013). After a restricted MIP is defined, it is typically solved using a commercial solver for no more than a few minutes, since limiting the solve time of restricted MIPs allows for more MIPs to be solved within a total time limit (Hwang *et al.* 2011, Erera *et al.* 2013). Similar to previous IP-based heuristics built for trucking network problems, our search neighborhoods are defined after selecting a small number of network components, such as locations and arcs, and selecting variables associated with these components. However, in contrast to previous work, our proposed IP-based heuristic identifies solution improvement opportunities via randomized search neighborhoods, where probabilities of selecting network components are computed to bias the search toward areas of the network with high freight volumes. We also use multiple search neighborhood generation schemes, and switch between them adaptively to escape local minima.

3. Middle-Mile Consolidation Optimization Modeling

In this section, we introduce the middle-mile consolidation network design problem with fixed origins and destinations. We first formulate a path-based mixed integer programming model using flat networks when lead-time requirements can be embedded in the set of possible consolidation routes. We then extend this base model and propose new flat network design MIPs that estimate waiting delays between load dispatches and ensure that shipment lead-time requirements are satisfied with a defined probability.

3.1. Problem Description

We consider a large shipper that needs to move shipments from known *origins* (vendor or fulfillment center (FC) locations) to known *destinations*, each of which is a last-mile distribution (LMD) facility, within specified lead times. To do so, the shipper needs to plan sufficient freight transportation capacity between its facilities to satisfy shipment demand and lead-time constraints. This capacity will be provided by scheduling *loads*, where a load is a consolidated set of shipments to be dispatched. We assume that the shipper does not own nor lease the transportation equipment assets used to move loads and is then not responsible for balancing these assets across facilities over time.

Let $(\mathcal{N}, \mathcal{L})$ define the shipper's service network. The node set $\mathcal{N}$ denotes the set of facilities in the network; these include vendor locations, FCs, LMD facilities, and potentially other sorting and transfer locations. Subset $\mathcal{N}_O \subseteq \mathcal{N}$ includes all locations that originate middle-mile shipments. Subset $\mathcal{N}_D \subseteq \mathcal{N}$ includes all locations that are destinations for shipments. Finally, $\mathcal{N}_H \subseteq \mathcal{N}$ includes all facilities where shipments can be transferred from one load to another; these intermediate locations may be FCs, cross-docks, or other transfer terminals, and we assume that they have sufficient capacity to transfer assigned shipments. Each facility $i \in \mathcal{N}$ belongs to at least one of the subsets $\mathcal{N}_O$, $\mathcal{N}_D$, or $\mathcal{N}_H$. The directed arc set $\mathcal{L}$ consists of the set of potential freight transportation legs connecting pairs of locations.

If shipments are moved on a leg $l \in \mathcal{L}$, they all must be assigned to a single *mode* $m \in \mathcal{M}_l$, and a leg-mode combination (l, m) will be referred to as a *lane*. Here, a mode $m \in \mathcal{M}_l$ indicates the type of freight transportation used on leg $l \in \mathcal{L}$ and also specifies cost parameters and bounds on the size of each individual load. For middle-mile networks, typical transportation types include truckload and LTL trucking. Given mode m, we assume that each load of size q dispatched on lane (l, m) incurs a cost given by the expression $A_{lm} + B_{lm}q$. This fixed-plus-linear form is a useful model that can represent many real-world freight cost structures reasonably well. Furthermore, each lane also specifies an associated upper bound Q_{lm}^{max} and lower bound Q_{lm}^{min} on the size of each dispatched load; these bounds are used to model both physical constraints on load size by mode and key size buckets where cost parameters differ. For example, a truckload mode can be modeled with a load size lower bound of zero and an upper bound equal to the maximum trailer capacity. On the other hand, an LTL mode may also specify a minimum load size required to qualify for a price discount. We restrict each leg to use a single mode to attempt to pragmatically represent operational realities; it is unlikely for a shipper to combine truckload and LTL shipments along a single leg, and while LTL shipments over time may vary in size (and thus size bucket), such variation is not important to capture in a planning model.

Given this network and available freight transportation modes, the middle-mile consolidation network design problem is to determine a minimum-cost allocation of transportation capacity on network legs to ensure that a shipment consolidation plan is feasible. Shipment demand is modeled using a set $\mathcal{K}$ of *commodities*. Since customer orders are filled from known origins to known destinations, each commodity $k \in \mathcal{K}$ has a fixed origin $o_k \in \mathcal{N}_O$ and destination $d_k \in \mathcal{N}_D$. Although many shipments may be sent over time for commodity k, we assume that each such shipment follows the same sequence (or route) defined by the chosen consolidation plan. Let $\mathcal{R}_k$ represent the set of potential freight routes for commodity k, where each route is an ordering of adjacent freight transportation legs connecting origin o_k to destination d_k, and potentially uses one or more transfer facilities in $\mathcal{N}_H$. Then, for each commodity $k \in \mathcal{K}$, a unique freight route $r \in \mathcal{R}_k$ must be selected. The selected route specifies a consolidation plan for commodity k: one that includes a single leg is referred to as a direct route, whereas a consolidation route has multiple legs and includes shipment transfer(s). For notational convenience, we denote $\mathcal{R} := \cup_{k \in \mathcal{K}} \mathcal{R}_k$ as the set of potential freight routes.

3.2. A Base Model of Middle-Mile Consolidation

We now introduce a base optimization model for middle-mile consolidation network design (MMC), which handles cases where shipment lead times can be determined *completely* by the legs and transfer terminals contained within each route. The proposed model uses a flat (not time-expanded) network representation of capacity allocation to legs and an associated representation of shipment consolidation into load dispatches. Thus, this base model is not a detailed schedule of actual planned load dispatches. Instead, freight transportation capacity decisions are modeled as frequencies of load dispatches on lanes per time (e.g., number of truckloads per week). The demand inputs then are also expressed as constant rates per time; let V_k be the demand rate for commodity k, representing the aggregated average shipment size flowing (i.e., the volume) from o_k to d_k per time (e.g., lbs per week). As a tactical model, it is assumed that any non-constant fluctuations in demand or load dispatch frequencies do not substantively impact the feasibility of the plan. The goal of the MMC model is to select a joint set of freight routes for all commodities along with load dispatch frequencies on selected lanes such that all commodity volume is transported feasibly and total cost is minimized.

Let binary variables x_r indicate whether route $r \in \mathcal{R}$ is selected and y_{lm} indicate whether lane $(l,m) \in \mathcal{L} \times \mathcal{M}_l$ is used. Continuous variables v_{lm} indicate the total shipment volume assigned to each lane (l,m). Finally, integer variables f_{lm} count the number (or frequency) of loads dispatched per time on lane (l,m). Suppose that shipment lead times can be *completely* determined by the legs and transfer terminals contained within each route. Furthermore, suppose that each route $r \in \mathcal{R}_k$ for commodity k has a total handling cost C_r, typically proportional to the number of transfers multiplied by the shipment volume V_k. Then, the set $\mathcal{R}$ can be pre-processed so that it only contains routes for which lead-time requirements are met, thus ensuring that the MMC model selects a consolidation plan that is lead-time feasible. We can formulate this model as follows:

$$
\begin{aligned}
\min_{x,y,f,v} \quad & \sum_{r \in \mathcal{R}} C_r x_r + \sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}_l} \left(A_{lm} f_{lm} + B_{lm} v_{lm} \right) && && \text{(1a)} \\
\text{s.t.} \quad & \sum_{r \in \mathcal{R}_k} x_r = 1, && \forall k \in \mathcal{K}, && \text{(1b)} \\
& \sum_{m \in \mathcal{M}_l} v_{lm} = \sum_{k \in \mathcal{K}} \sum_{\{r \in \mathcal{R}_k | r \ni l\}} V_k x_r, && \forall l \in \mathcal{L}, && \text{(1c)} \\
& v_{lm} \le Q^{max}_{lm} f_{lm}, && \forall l \in \mathcal{L},\ \forall m \in \mathcal{M}_l, && \text{(1d)} \\
& v_{lm} \ge Q^{min}_{lm} f_{lm}, && \forall l \in \mathcal{L},\ \forall m \in \mathcal{M}_l, && \text{(1e)} \\
& \sum_{m \in \mathcal{M}_l} y_{lm} \le 1, && \forall l \in \mathcal{L}, && \text{(1f)} \\
& f_{lm} \le F_{lm} y_{lm}, && \forall l \in \mathcal{L},\ \forall m \in \mathcal{M}_l, && \text{(1g)} \\
& x_r \in \{0,1\}, && \forall r \in \mathcal{R}, && \text{(1h)} \\
& y_{lm} \in \{0,1\}, && \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l, && \text{(1i)} \\
& f_{lm} \in \mathbb{Z}_{\ge 0}, && \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l, && \text{(1j)} \\
& v_{lm} \ge 0, && \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l. && \text{(1k)}
\end{aligned}
$$

The objective is to determine a transportation consolidation plan that minimizes the total transportation and handling costs. Constraints (1b) ensure that one route is selected for each commodity. Constraints (1c) determine the total volume flowing on each leg l aggregated across commodities and allocate it to a selected lane (l,m). Constraints (1d) and (1e) set the required load dispatch frequencies for each lane using upper and lower bounds on load size. Note that, consistent with nearly all flow and load planning models in the literature, these constraints assume that shipments can be fluidly packed into loads ignoring discrete bin packing considerations. Constraints (1f) ensure that each leg uses at most one mode (and thus is included in at most one lane). Finally, since the number of dispatched loads using lane (l,m) may be limited over time (especially for LTL shipments), constraints (1g) require that a lane-specific maximum load dispatch frequency F_{lm} is not exceeded. Note also that this model does not include constraints to balance transport equipment across locations over time, as the shipper does not own the transportation equipment assets.

The MMC model is formulated using path-based variables x_r since in most middle-mile planning problems, the number of reasonable geographic consolidation routes per commodity is likely to be small when compared to the number of network arcs. In such scenarios, path-based models require fewer binary decision variables. It is also well-known that path-based models make it very easy to exclude certain routes from consideration; for example, routes that induce more than some allowable fraction of out-of-route mileage or that require more than some maximum transit time can be excluded easily. In Section 3.4, it will also become clear that path-based models are useful when modeling probabilistic commodity lead-time constraints.

A key concern about the MMC model is its ability to capture true shipment lead times since it ignores *waiting delays* for load dispatches. If the solution has sufficiently high frequencies on all lanes (i.e., dispatches daily or more frequently), then it may be appropriate to ignore waiting delays. However, when loads are potentially dispatched less frequently, as is typically the case with large-and-bulky items for instance, it becomes crucial to explicitly model waiting delays. In the subsequent sections, we develop approaches for doing so.

3.3. Modeling Lead-Time Constraints with Waiting Time Delays

Missing from the MMC base model is a representation of the waiting delays that shipments experience given a consolidation plan. If the shipments for a commodity travel along a route r with a single leg (and associated lane), they incur potential waiting delays for loads departing the origin. When shipments additionally transfer along a route with multiple legs, waiting delay may occur at each dispatch location.

The frequency of load dispatches on a leg impacts lead times since lower frequencies lead to longer waiting delays. Given the load dispatch frequency f_l on leg l, we assume trucks are scheduled to dispatch every $\frac{1}{f_l}$ time units, which we refer to as the resulting *headway*. We can then safely make the simplifying assumption that load dispatches and headways are deterministic and uncoordinated across facilities. Then, if individual shipment sizes are small compared to the capacity of each load and they are equally likely to arrive at any time within a dispatch headway of length $\frac{1}{f_l}$, the waiting delay to the end of the headway (and subsequent dispatch) experienced by any individual shipment to be dispatched on leg l can be modeled as a uniform random variable $W_l \sim \text{Uniform}(0, \frac{1}{f_l})$. Since the flat network model assumes that originating shipment volume arrives continuously and deterministically over time at origin locations, it should be clear that the waiting delay for the initial dispatch is uniformly distributed. When volume is transferred at an intermediate location, if inbound arriving loads and headways are uncoordinated, it is again most reasonable to assume that any individual shipment's arrival time is uniformly-distributed on the headway interval and that the shipment can be loaded into the next dispatch. Now suppose that we replace the fluid arrival model of shipments at each origin with a homogeneous Poisson process; again, the waiting time from any individual shipment's arrival time to the next dispatch will be $\text{Uniform}(0, \frac{1}{f_l})$, since the arrival times of an observed set of Poisson points on an interval of known length are uniformly distributed on that interval. When individual shipment sizes are small compared to the size of a consolidated load, the uniform waiting delay model is again appropriate.

We now define the lead time of a route as the sum of leg transit times and waiting delays for load dispatches, where we assume that any shipment processing time at an intermediate facility $h \in \mathcal{N}_H$ is included in the transit time of the outbound leg and is independent of the total volume that moves through the transfer location. In Figure 1, we illustrate how commodity arrival times and uncoordinated dispatches can lead to different waiting delays experienced by two shipments of the same commodity. In the example, the shipments must traverse a two-leg route (with fixed transit times of 0.5 days and 1 day, respectively) to reach their destination. Since dispatches are not coordinated, we observe that the first shipment incurs 2 days of waiting delay, while the second incurs 2.6 days of waiting delay.

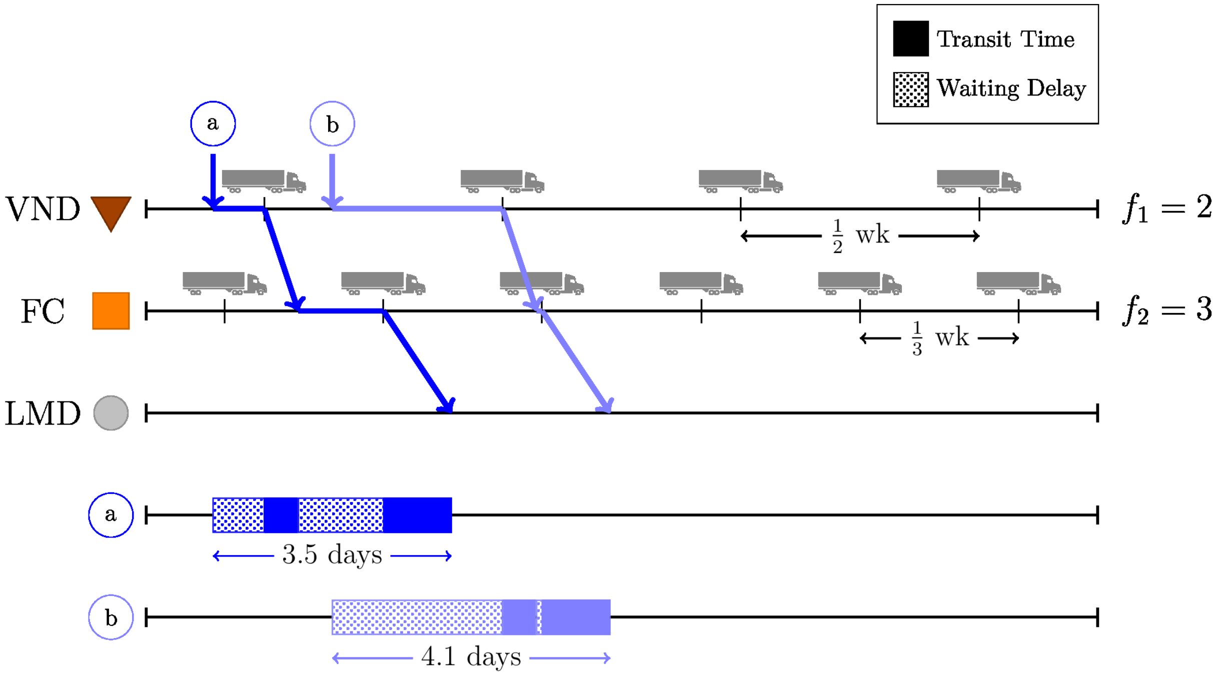


Figure 1 **Lead time illustration of two shipments of a single commodity.**

The allowable waiting delay of route r, denoted $\hat{W}_r$, is its lead-time requirement less the sum of its leg transit times. A load plan will then satisfy the lead-time requirement of route r if and only if the total waiting delay along that route does not exceed $\hat{W}_r$. This can be expressed as

$$\sum_{l \in r} W_l \leq \hat{W}_r, \tag{2}$$

which involves the random variables W_l. To construct a plan that meets customer service goals, one could choose from several potential models. For example, a robust plan might seek to satisfy (2) with probability one. One could also compute expected plan lateness by taking the expectation of the maximum of the left-hand side less the right-hand side and zero for each commodity, and then summing these commodity lateness expressions weighted by volume. Expected plan lateness could be penalized in the objective function or treated as a second objective function to cost in a biobjective model.

Instead, we choose to develop a general chance-constrained model that ensures (under assumptions) that any shipment is on time with a certain probability p. Consider then the following chance constraint:

$$\mathbb{P}\left(\sum_{l \in r} W_l \leq \hat{W}_r\right) \geq p. \tag{3}$$

Constraint (3) ensures that the probability the sum of the dispatch waiting delays does not exceed the allowable waiting delay is at least p. Specifically, p represents the probability guarantee of an on-time arrival for the commodity using route r and is selected by the shipper as a lower bound on service quality guaranteed to the customer. Since the waiting delay experienced by shipments to be dispatched on leg l is assumed to be given by $W_l \sim \text{Uniform}\left(0, \frac{1}{f_l}\right)$, the probability that the commodity traveling along r arrives on time to its destination is given by the following expression (Kang *et al.* 2010):

$$\mathbb{P}\left(\sum_{l \in r} W_l \leq \hat{W}_r\right) = \frac{1}{|r|! \prod_{l \in r} \frac{1}{f_l}} \sum_{J \subseteq r} (-1)^{|J|} \left[\max\left\{0, \hat{W}_r - \sum_{l \in J} \frac{1}{f_l}\right\}\right]^{|r|}. \tag{4}$$

However, the resulting constraint (3) is nonlinear in the load dispatch frequencies and cannot be included directly in the optimization model (1). Instead, we approximate (3) using integer linear constraints and add them to model (1).

The first step consists of approximating constraint (3) using a simpler nonlinear constraint by making the following observations: When $p = 0.5$, i.e., the probability that the commodity traveling on route r arrives on time is 0.5, chance constraint (3) is equivalent to:

$$\sum_{l \in r} \frac{1}{2} \frac{1}{f_l} \leq \hat{W}_r\,. \tag{5}$$

Similarly, when the probability of on-time arrival is $p = 1$, chance constraint (3) is equivalent to:

$$\sum_{l \in r} \frac{1}{f_l} \leq \hat{W}_r \,. \tag{6}$$

From these observations, given a general on-time arrival probability p, we approximate chance constraint (3) using the following constraint:

$$\rho_r \sum_{l \in r} \frac{1}{f_l} \leq \hat{W}_r \,, \tag{7}$$

where $\rho_r \in [0, 1]$ is a conservatism hyperparameter for route r that depends on its allowable waiting delay $\hat{W}_r$ and number of legs, as well as the probability guarantee p. Put simply, the selected ρ_r represents the maximum proportion of the headway (i.e., $\sum_{l \in r} \frac{1}{f_l}$) the commodity can wait in order to guarantee the probability p of meeting the commodity lead-time requirement desired by the shipper.

Given deterministic and homogeneous headways, setting $\rho_r = 0.5$ (respectively $\rho_r = 1$) ensures that feasible load plans satisfy the lead-time requirement for route r with probability $p = 0.5$ (respectively $p = 1$). However, in general, determining the hyperparameters ρ_r given the desired on-time arrival probability p is challenging. A low value of ρ_r will allow the selection of load plans that will not meet the on-time arrival probability, while a high value of ρ_r will force the selection of load plans that are too conservative and costly. Thus, given a probability p of on-time arrival desired by the shipper, we consider the problem of determining for each route r the lowest hyperparameter ρ_r for which constraints (7) guarantee that a commodity traveling on route r meets the lead-time requirement with probability p. This problem can be formulated as follows:

$$\begin{aligned} &\min \ \rho_r = \frac{\hat{W}_r}{\sum_{l \in r} \frac{1}{f'_l}} \\ &\text{s.t.} \ \mathbb{P}\left(\sum_{l \in r} W_l \leq \hat{W}_r\right) \geq p, \quad \forall\, f \in \mathbb{Z}^r_{>0} \mid \rho_r \sum_{l \in r} \frac{1}{f_l} \leq \hat{W}_r, \\ &\quad f' \in \mathbb{Z}^r_{>0}. \end{aligned}$$

Equivalently, we select for each route r the lowest value of ρ_r that excludes any combination of load dispatch frequencies with a total waiting delay that exceeds the allowable waiting delay with probability at least $1 - p$. Whenever the number of legs per route is small, which is typically the case in consolidation transportation systems, this problem can be solved by smartly iterating over permissible load dispatch frequencies until we are guaranteed that the remaining load plans satisfy the on-time probability. Note that in some cases, there does not exist a ρ_r value that separates all combinations of load dispatch frequencies that satisfy chance constraint (3) from the ones that do not. Thus, this approach can lead to setting ρ_r to a value that will exclude a small number of load dispatch frequency combinations that result in an on-time probability at least p and thus may be more conservative than necessary.

Although the non-linear constraints (7) are a rather simple sum of separable hyperbolic terms for each route, incorporating constraints of this type directly into the MMC model is not straightforward. In the next section, we discuss our approach to reformulate the MMC model in such a way that allows us to linearize the lead-time constraints.

3.4. A Middle-Mile Consolidation Model with Linearized Lead-Time Constraints

To include lead-time constraints in the MMC model, we reformulate constraints (7) using binary variables. We call the resulting optimization model the *middle-mile consolidation with waiting times* (MMCW) model. The linearization approach is similar to that proposed in Cancela *et al.* (2015) for transit network design problems. For each lane (l, m) and each possible positive frequency $\omega \in \mathcal{F}_{lm} := \{1, \ldots, F_{lm}\}$ satisfying the maximum load dispatch frequency F_{lm}, we define the binary variable $z_{lm\omega}$. We then substitute the frequency variables as follows:

$$f_{lm} = \sum_{\omega \in \mathcal{F}_{lm}} \omega z_{lm\omega}, \quad \forall\, l \in \mathcal{L},\ \forall\, m \in \mathcal{M}_l\,. \tag{8}$$

We can now reformulate (7) as linear integer constraints, thus leading to the following formulation of the MMCW model:

$$\min_{x,z,v} \quad \sum_{r \in \mathcal{R}} C_r x_r + \sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}_l} \Big[A_{lm}\Big(\sum_{\omega \in \mathcal{F}_{lm}} \omega z_{lm\omega}\Big) + B_{lm} v_{lm}\Big] \tag{9a}$$

$$\text{s.t.} \quad \sum_{r \in \mathcal{R}_k} x_r = 1, \qquad \forall\, k \in \mathcal{K}, \tag{9b}$$

$$\sum_{m \in \mathcal{M}_l} v_{lm} = \sum_{k \in \mathcal{K}} \sum_{\{r \in \mathcal{R}_k | r \ni l\}} V_k x_r, \qquad \forall\, l \in \mathcal{L}, \tag{9c}$$

$$v_{lm} \leq Q^{max}_{lm} \sum_{\omega \in \mathcal{F}_{lm}} \omega z_{lm\omega}, \qquad \forall\, l \in \mathcal{L},\ \forall\, m \in \mathcal{M}_l, \tag{9d}$$

$$v_{lm} \geq Q^{min}_{lm} \sum_{\omega \in \mathcal{F}_{lm}} \omega z_{lm\omega}, \qquad \forall\, l \in \mathcal{L},\ \forall\, m \in \mathcal{M}_l, \tag{9e}$$

$$\sum_{m \in \mathcal{M}_l} \sum_{\omega \in \mathcal{F}_{lm}} z_{lm\omega} \leq 1, \qquad \forall\, l \in \mathcal{L}, \tag{9f}$$

$$\rho_r \sum_{l \in r} \sum_{m \in \mathcal{M}_l} \sum_{\omega \in \mathcal{F}_{lm}} \frac{1}{\omega} z_{lm\omega} \leq \hat{W}_r x_r + \rho_r |r| (1 - x_r), \qquad \forall\, r \in \mathcal{R}, \tag{9g}$$

$$x_r \in \{0, 1\}, \qquad \forall\, r \in \mathcal{R}, \tag{9h}$$

$$z_{lm\omega} \in \{0, 1\}, \qquad \forall\, l \in \mathcal{L}, \forall\, m \in \mathcal{M}_l, \forall\, \omega \in \mathcal{F}_{lm}, \tag{9i}$$

$$v_{lm} \geq 0, \qquad \forall\, l \in \mathcal{L}, \forall\, m \in \mathcal{M}_l. \tag{9j}$$

Objective (9a) and constraints (9d)-(9e) are obtained by applying the variable replacement (8) to objective (1a) and constraints (1d)-(1e), respectively. Constraints (9f) replace constraints (1f)-(1g) and select at most one frequency for each lane. Constraints (9g) provide a linear formulation of constraints (7) and ensure that commodities arrive on time to their destinations with probability at least p. Note that if route r is not selected and $x_r = 0$, the second term on the right-hand side provides an upper bound on the left-hand side, which is largest when all lanes in the route are given the minimum non-zero frequency values of 1. Finally, notice that the MMCW model does not need binary variables y_{lm} or frequency upper bound constraints since they are accounted for in the sets $\mathcal{F}_{lm}$ of allowed frequency values.

3.5. Simplifying Lead-Time Constraints by Allocating Allowable Wait

Note that the main drawback of the MMCW model is the potentially large number of binary variables $z_{lm\omega}$, $\forall (l,m,\omega) \in \mathcal{L} \times \mathcal{M}_l \times \mathcal{F}_{lm}$ needed when lanes have many possible frequency values; this issue is exacerbated for large-scale networks with many legs, lanes, and commodities. To manage this challenge, one could restrict the allowable frequency values to very small cardinality sets for a large set of lanes. For example, if it is likely that a lane (l,m) will be used to transport a large shipment volume, the minimum load dispatch frequency can be increased to reduce the cardinality of $\mathcal{F}_{lm}$. Of course, such a restriction approach requires insight on potential solutions and leads to an upper bound on the optimal solution value to the MMCW problem.

As an alternative, we develop a restricted version of the MMCW problem (and an associated upper bound on its optimal value) that greatly reduces the size of the MIP formulation and can be used to produce very good starting solutions for the MMCW problem when solving large, real-world instances. This approach restricts the space of feasible solutions by allocating fixed fractions of a route's total allowable waiting delay *a priori* to each of its legs, and by doing so we can build waiting time constraints directly in the space of the original decision variables of the MMC model. We denote this restricted model as the *middle-mile consolidation with allocated waiting delay* (MMCW-A) model. In general, this allocation can be arbitrary with the only constraint that the sum of the individual leg allowable delays for route r does not exceed $\hat{W}_r$. For this paper, however, we limit our attention to a simple strategy that distributes the total allowable delay equally among the legs of each route. Under this assumption, we now approximate chance constraint (3) using the following hyperparameterized constraint on every leg l of a selected route r:

$$\rho_r \sum_{m \in \mathcal{M}_l} \frac{1}{f_{lm}} \leq \frac{\hat{W}_r}{|r|}.$$

This is equivalent to directly adding the following linear constraints to (1) to yield the MMCW-A model:

$$\sum_{m \in \mathcal{M}_l} f_{lm} \geq \rho_r \frac{|r|}{\hat{W}_r} x_r, \quad \forall r \in \mathcal{R},\ \forall l \in r. \tag{10}$$

The right-hand side of constraints (10) represents the minimum frequency of load dispatches on each leg of a route that is needed to ensure the desired on-time arrival probability of a commodity traveling on route r, for an appropriately selected hyperparameter ρ_r. Note that we are able to rearrange the terms in this manner because only one load dispatch frequency variable for every leg l will be non-zero, given constraints (1f) and (1g). Importantly and in contrast to the MMCW model, the MMCW-A model does not require additional binary variables (although it retains the integer lane frequency and binary lane variables). In practice, it is likely to be much simpler computationally to find feasible and optimal solutions to this model. Of course, allocating allowable waiting delays *a priori* may lead to suboptimal solutions. This may occur, for example, when a large shipment volume is assigned to a leg l_1, leading the capacity constraint (1d) to set a high load dispatch frequency, which in turn may result in a waiting delay significantly lower than the allowable waiting delay allocated to l_1. If a consolidation route contains leg l_1 and a leg l_2 with low assigned volume, then the load dispatch frequency for l_2 needed to meet the route's lead-time constraint could be lower than what was permissible by constraint (10) and the allowable waiting delay allocated to l_2.

Similar to the MMCW model, we aim to determine for every selected route $r \in \mathcal{R}$ the lowest hyperparameter ρ_r that will guarantee that the load plans satisfying constraints (10) meet the corresponding lead-time requirement with probability at least p. This problem can be formulated as follows:

$$\begin{aligned}
&\min\ \rho_r = \frac{f'\hat{W}_r}{|r|} \\
&\text{s.t. } \mathbb{P}\left(\sum_{l\in r} W_l \le \hat{W}_r\right) \ge p, \quad \forall f \in \mathbb{Z}^r_{>0} \mid \frac{1}{f_l} \le \frac{\hat{W}_r}{\rho_r |r|}\ \forall l \in r, \\
&\qquad f' \in \mathbb{Z}_{>0}.
\end{aligned}$$

Interestingly, we can derive the following result for the load dispatch frequencies satisfying constraints (10) (see Appendix A for the complete derivation). Specifically, for every route $r \in \mathcal{R}$ and every set of frequencies $f \in \mathbb{Z}^r_{\ge 0}$ satisfying $\frac{1}{f_l} \le \frac{\hat{W}_r}{\rho_r|r|}$ for every $l \in r$, we have:

$$\mathbb{P}\left(\sum_{l\in r} W_l \le \hat{W}_r\right) \ge \sum_{i=0}^{\lfloor \rho_r |r| \rfloor} \frac{(-1)^i}{i!(|r|-i)!} (\rho_r|r| - i)^{|r|} =: g_r(\rho_r). \tag{11}$$

Thus, given a desired on-time probability p, we can determine the corresponding conservatism level ρ_r for each constraint (10) using an iterative search (e.g., bisection), since g_r is a nondecreasing function. Alternatively, if the number of legs $|r|$ is small, one can determine the conservatism level ρ_r by solving the polynomial equation $g_r(x) - p = 0$ on each interval $[\frac{j-1}{|r|}, \frac{j}{|r|}]$, $j \in \{1, \ldots, |r|\}$.

Surprisingly, we observe that computing the conservatism level ρ_r for the MMCW-A approach simply requires the on-time probability p and number of legs $|r|$, and is independent of the allowed waiting delay $\hat{W}_r$. Figure 2 illustrates the guaranteed on-time probability as a function of the conservatism level and the number of legs for this model.

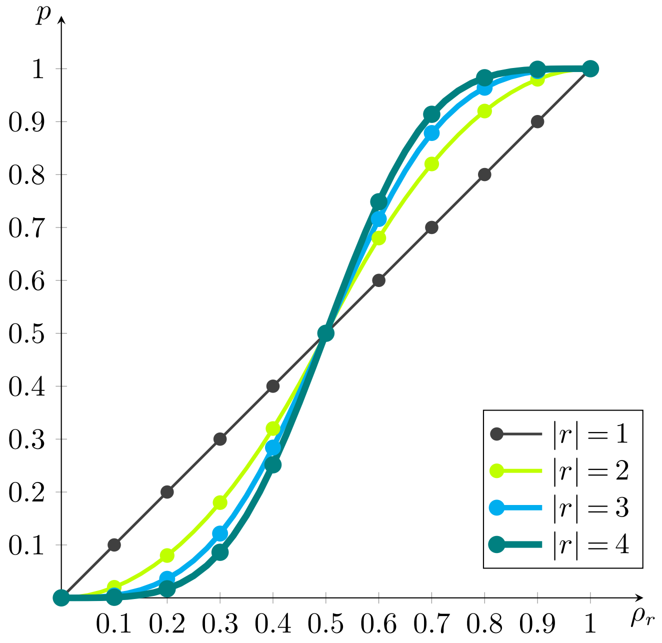


Figure 2 Guaranteed probability p of on-time arrival given homogeneous conservatism levels ρ_r using the MMCW-A model.

Again, we find that when the probability of on-time arrival is $p = 0.5$ (resp. $p = 1$), setting the hyperparameter to $\rho_r = 0.5$ (resp. $\rho_r = 1$) ensures that the load plans that satisfy constraints (10) are guaranteed to meet the lead-time requirement for the commodity traveling on route r with probability p, regardless of the number of legs $|r|$. However, it is interesting to note that if $\rho_r > 0.5$, the probability of on-time arrival increases with the number of legs $|r|$. On the other hand, if $\rho_r < 0.5$, the probability of on-time arrival decreases with $|r|$. This is a consequence of the independence between the waiting delay distributions within a route. This observation suggests that there is value in adapting the conservatism level for each route, rather than selecting a unique conservatism level for all routes.

We analyze the computational benefits and drawbacks of both the MMCW and MMCW-A models, as well as the effects of varying conservatism levels, later in Section 5. Notably, since a solution to the MMCW-A optimization problem is always feasible for the MMCW problem, we will show how to use both models in tandem in a two-phase heuristic solution approach useful for the largest problem instances.

4. IP-Based Local Search Heuristic

Solving all models proposed in this paper for large-scale realistic problem instances is very challenging. Commercial solvers often fail to obtain feasible solutions to larger instances with a reasonable optimality gap, and almost never find provably optimal solutions. Compared to the MMC model, the MMCW model includes only $|\mathcal{R}|$ additional constraints to model the lead-time upper bound for each possible route and removes $\sum_{l\in\mathcal{L}} |\mathcal{M}_l|$ integer variables. However, it requires a very large number of additional binary variables ($\sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}_l} (|\mathcal{F}_{lm}|-1)$) and the resulting MIP proves much harder to solve than MMC instances for identical networks. Similarly, instances of the MMCW-A model are challenging to solve due to the large number of allowable waiting delay constraints added, which is equal to $\sum_{r\in\mathcal{R}} |r|$.

For these reasons, we develop IP-based local search (IPBLS) solution approaches (see Erera *et al.*, 2013 and Lindsey *et al.*, 2016) to effectively find good solutions to all three models. These approaches work by iteratively improving an incumbent feasible solution by solving small, restricted versions of the MIP models proposed earlier. These restricted models are obtained by fixing most decision variables to their values in the incumbent, and then optimizing over the remaining variables. The incumbent always remains feasible and is passed to the solver as a warm-start solution. Then, the restricted MIP is solved using a time limit to generate a new incumbent; note that the warm-start solution in some cases may not be improved. The search continues by defining and solving a new restricted MIP at each iteration until a stopping criterion is met.

To ensure that the IPBLS solution approach is effective, we must determine which set of variables to optimize over at each iteration. Each restricted optimization can be considered as the selection of a (potentially) improving solution in a randomized neighborhood defined by the free decision variables. We use two approaches to generate such neighborhoods, and each is motivated by the premise that locations with large originating shipment volume often drive consolidation decisions. Once consolidation legs are used in paths for origins with larger volumes, these legs become attractive in paths for vendors with smaller volumes as well. The first generated neighborhood, Neighborhood 1, seeks to improve consolidation outbound from origin facilities (vendors or FCs), biased toward origins that have more outbound demand volume. The second generated neighborhood, Neighborhood 2, is similar but focuses on origin facilities that might utilize a particular intermediate terminal as a consolidation point, again biased toward those with more demand volume.

We now describe in detail how we generate the IPBLS neighborhoods. Given an incumbent solution, a randomized neighborhood of feasible solutions is generated by fixing all the route decision variables x_r associated with a subset of commodities, while freeing up all other decision variables. At iteration t of the IPBLS, let $\mathcal{R}^{(t)}$ be the set of routes whose decision variables x_r will be freed for reoptimization. Algorithm 1 specifies how such a subset $\mathcal{R}^{(t)}$ is selected to define a randomized instance of Neighborhood 1. Each iteration, an origin facility o_s is selected and all routes in $\mathcal{R}_k$ for all commodities k that originate at o_s are added to $\mathcal{R}^{(t)}$. The origin o_s is selected at random each iteration, where the probability of selecting an origin is equal to its fraction of the total outbound shipment volume remaining among all origin facilities o that have not yet been selected. This process continues with another iteration until at least $\alpha|\mathcal{R}|$ routes are included in $\mathcal{R}^{(t)}$, where $\alpha \in (0,1]$ is a user-selected parameter. A randomized instance of Neighborhood 2 is generated using a similar procedure, as specified in Algorithm 2. Again, an origin facility o_s is identified at each iteration and all routes for all commodities originating at o_s are added to $\mathcal{R}^{(t)}$. However, Algorithm 2 selects o_s from a subset of origin facilities $\mathcal{O}_h$ that can transport some of their outbound volume through a specific intermediate transfer facility $h \in \mathcal{N}_H$. Here, the probability of selecting a specific origin $o_s \in \mathcal{O}_h$ is given by its fraction of outbound shipment volume in commodities that have a route including transfer location h among all the remaining origin facilities in $\mathcal{O}_h$. Note here that *all* commodities originating at o_s and their associated routes are freed for reoptimization. Doing so provides the flexibility for other commodities to shift from direct routes to consolidation routes, or vice versa, as the majority of outbound volume from an origin should often flow altogether to or be removed altogether from a common initial consolidation location to be cost-effective.

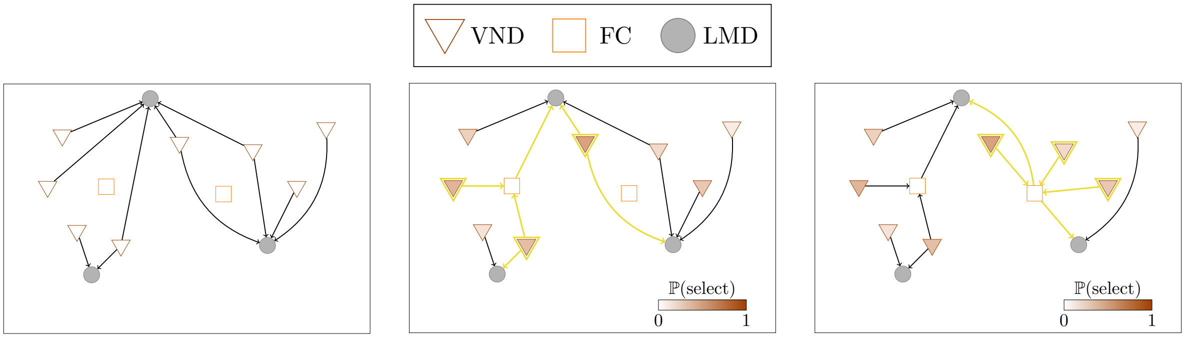
Algorithm 1: Route Set $\mathcal{R}^{(t)}$ Selection for IPBLS Neighborhood 1

```
Input: Route set $\mathcal{R}$, commodity set $\mathcal{K}$, commodity volumes $V_k,\ \forall k \in \mathcal{K}$, percentage of
       routes to add $\alpha$
Result: Route subset $\mathcal{R}^{(t)}$
Set $\mathcal{R}^{(t)} \leftarrow \emptyset$;
Set $\mathcal{O} \leftarrow \{o_k,\ \forall k \in \mathcal{K}\}$;
Set $\hat{V} \leftarrow \sum_{k \in \mathcal{K}} V_k$;
while $|\mathcal{R}^{(t)}| < \alpha|\mathcal{R}|$ and $\mathcal{O} \neq \emptyset$ do
    Set $w(o) \leftarrow \frac{1}{\hat{V}} \sum_{\{k \in \mathcal{K} \,|\, o_k = o\}} V_k,\ \forall o \in \mathcal{O}$;
    Select origin $o_s$ randomly from $\mathcal{O}$ using probability mass function $w$;
    $\mathcal{R}^{(t)} \leftarrow \mathcal{R}^{(t)} \cup \left(\cup_{\{k \in \mathcal{K} \,|\, o_k = o_s\}} \mathcal{R}_k\right)$;
    $\mathcal{O} \leftarrow \mathcal{O} \setminus \{o_s\}$;
    $\hat{V} \leftarrow \hat{V} - \sum_{\{k \in \mathcal{K} \,|\, o_k = o_s\}} V_k$;
end
return $\mathcal{R}^{(t)}$
```

Algorithm 2: Route Set $\mathcal{R}^{(t)}$ Selection for IPBLS Neighborhood 2

```
Input: Route set $\mathcal{R}$, commodity set $\mathcal{K}$, commodity volumes $V_k,\ \forall k \in \mathcal{K}$, selected
       intermediate facility $h \in \mathcal{N}_H$, percentage of routes to add $\alpha$
Result: Route subset $\mathcal{R}^{(t)}$
Set $\mathcal{R}^{(t)} \leftarrow \emptyset$;
Set $\mathcal{K}_h \leftarrow \{k \in \mathcal{K} \mid$ at least one route $r \in \mathcal{R}_k$ includes location $h$ as a transfer point$\}$;
Set $\mathcal{O}_h \leftarrow \{o_k,\ \forall k \in \mathcal{K}_h\}$;
Set $\hat{V} \leftarrow \sum_{k \in \mathcal{K}_h} V_k$;
while $|\mathcal{R}^{(t)}| < \alpha |\mathcal{R}|$ and $\mathcal{O}_h \neq \emptyset$ do
    Set $w(o) \leftarrow \frac{1}{\hat{V}} \sum_{\{k \in \mathcal{K}_h \mid o_k = o\}} V_k,\ \forall o \in \mathcal{O}_h$;
    Select origin $o_s$ randomly from $\mathcal{O}$ using probability mass function $w$;
    $\mathcal{R}^{(t)} \leftarrow \mathcal{R}^{(t)} \cup \left(\cup_{\{k \in \mathcal{K}_h \mid o_k = o_s\}} \mathcal{R}_k\right)$;
    $\mathcal{O}_h \leftarrow \mathcal{O}_h \setminus \{o_s\}$;
    $\hat{V} \leftarrow \hat{V} - \sum_{\{k \in \mathcal{K}_h \mid o_k = o_s\}} V_k$;
end
return $\mathcal{R}^{(t)}$
```

Given these neighborhood generation methods, the IPBLS proceeds as detailed in Algorithm 3. First, an initial feasible solution is created as input to the search. Typically, each commodity will have a single-leg direct route that can be selected and which will result in a feasible solution as long as the frequency upper bounds F_{lm} are not restrictive; we will not focus in this paper on finding a good initial feasible solution in general. Next, the IPBLS begins by using randomized Neighborhood 1. When a solution is found that improves the objective value of the incumbent, the incumbent is updated. Within each iteration of the search, the incumbent solution is used as a warm-start solution. Simple illustrative examples of the IPBLS when using Neighborhoods 1 and 2 are given in Figures 3 and 4, respectively. In each figure, a initial feasible solution is shown in subfigure (a) and subfigures (b) and (c) illustrate two iterations of the IPBLS. Specifically, in Figure 3, a subset of origins is selected at random, where the probability of selecting ($\mathbb{P}$(select)) an origin is equal to its proportion of the total volume, and their commodity routes are optimized. In Figure 4, an FC is first selected and a subset of origins from those that have the option of transferring at the chosen FC are randomly selected, where the probability of selecting ($\mathbb{P}$(select)) an origin is equal to its proportion of the total volume that may flow through the selected FC. The commodity routes of the selected origins are then optimized. If the incumbent solution is not improved using randomized instances generated by the current neighborhood (1 or 2) for a number of consecutive iterations, the search switches to the other neighborhood. The search terminates once a time limit has been reached. Algorithm 3 is labeled as the Single-Phase IPBLS since it is used to solve an instance of either the MMC, MMCW, or MMCW-A model directly.

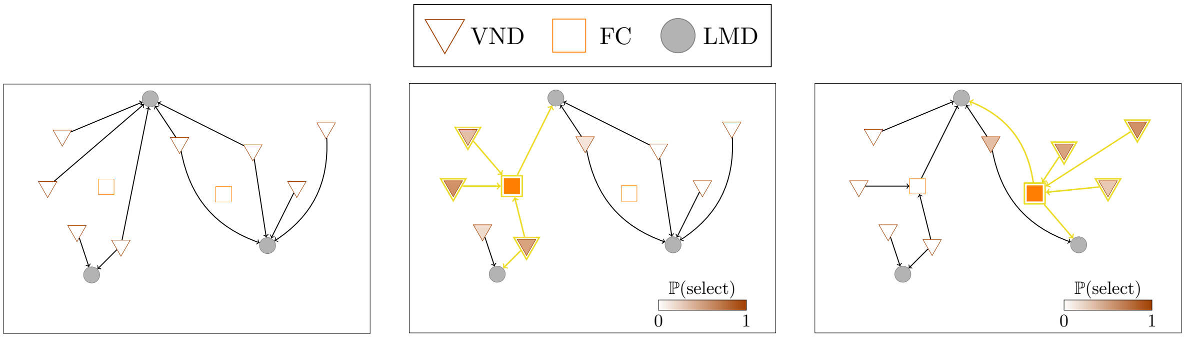


(a) Initial feasible solution.

(b) Origins assigned ℙ(select) according to proportion of total volume and randomly selected; commodity routes optimized.

(c) ℙ(select) values remain the same; next set of origins randomly selected; commodity routes optimized.

Figure 3 Illustration of IPBLS when using Neighborhood 1 to select route decision variables.



(a) Initial feasible solution.

(b) FC selected; origins assigned ℙ(select) according to portion of volume able to use selected FC; origins randomly selected; commodity routes optimized.

(c) Next FC selected; origins assigned ℙ(select) according to portion of volume able to use selected FC; origins randomly selected; commodity routes optimized.

Figure 4 Illustration of IPBLS when using Neighborhood 2 to select route decision variables.

Since large MMCW instances are particularly challenging to solve, we also develop a two-phase approach that leads to better solutions in faster solve times. In this approach, we take advantage of the fact that an MMCW-A model instance is a restriction of a corresponding MMCW instance. Thus, we can first improve an initial feasible solution to the restricted MMCW-A model instance using Algorithm 3. Once a time limit is reached, the feasible solution found is used as the new initial solution for a second run of Algorithm 3 using the MMCW instance to complete the solve. This two-phase IPBLS approach is detailed in Algorithm 4. Note that the total allowed run time T of this two-phase algorithm is allocated in advance to time $T_{\text{MMCW-A}}$ spent improving the solution using the restricted MMCW-A model and time T_{MMCW} spent improving the solution using the complete MMCW model.

Algorithm 3: Single-Phase IP-Based Local Search

```
Input: MIP, initial feasible solution $(\hat{x}, \hat{v}, (\hat{f}, \hat{y})$ or $\hat{z})$, objective value $\hat{w}$, and $list_H$ as an
       ordered list of transfer locations $\mathcal{N}_H$
Result: Improved feasible solution and improved objective value
Set $val \leftarrow \hat{w}$,   $T_{run} \leftarrow 0$,   $iter \leftarrow 0$,   $neighborhood\_select \leftarrow 1$,   and $i \leftarrow 1$;
while $T_{run} \leq T$ do
    if $neighborhood\_select = 1$ then
        Select $\mathcal{R}^{(t)}$ using Algorithm 1;
    else
        $h \leftarrow list_H[i]$;
        Select $\mathcal{R}^{(t)}$ using Algorithm 2 and $h$ as selected intermediate facility;
        if $i < |list_H|$ then
            $i \leftarrow i + 1$;
        else
            $i \leftarrow 1$;
    Solve MIP after adding constraints $x_r = \hat{x}_r$, $\forall r \in \mathcal{R} \backslash \mathcal{R}^{(t)}$, using $(\hat{x}, \hat{v}, (\hat{f}, \hat{y})$ or $\hat{z})$ as
     warm-start solution;
    $T_{\text{MIP}} \leftarrow$ MIP solving time;
    $newval \leftarrow$ MIP solution's objective value;
    if $newval < val$ then
        Set $(\hat{x}, \hat{v}, (\hat{f}, \hat{y})$ or $\hat{z}) \leftarrow$ MIP solution;
        Set $val \leftarrow newval$,   $iter \leftarrow 0$;
    else
        Set $iter \leftarrow iter + 1$;
    if $iter = N$ then
        if $neighborhood\_select = 1$ then
            $neighborhood\_select \leftarrow 2$;
        else
            $neighborhood\_select \leftarrow 1$;
        Set $iter \leftarrow 0$;
    $T_{run} \leftarrow T_{run} + T_{\text{MIP}}$;
end
return $(\hat{x}, \hat{v}, (\hat{f}, \hat{y})$ or $\hat{z})$, $val$
```

5. Computational Results

In this section, we describe the design and the results of a computational study to analyze the middle-mile consolidation plans produced by the models proposed in this paper and to evaluate the performance of our heuristic solution approaches. In particular, we present results that: (i) provide insights on the solution characteristics of the plans produced using MMCW models; (ii)

```
Algorithm 4: Two-Phase IP-Based Local Search for MMCW Model
Input: Initial feasible solution $(\hat{x}, \hat{v}, \hat{z})$ and objective value $\hat{w}$
Result: Improved feasible solution and improved objective value
Set $\hat{f}_{lm} \leftarrow \sum_{f \in \mathcal{F}_{lm}} f \hat{z}_{lmf} \quad \forall\, l \in \mathcal{L}, \forall\, m \in \mathcal{M}_l$;
Set $\hat{y}_{lm} \leftarrow \sum_{f \in \mathcal{F}_{lm}} \hat{z}_{lmf} \quad \forall\, l \in \mathcal{L}, \forall\, m \in \mathcal{M}_l$;
Set $val \leftarrow \hat{w}$;
Run single-phase IPBLS (Algorithm 3) using MMCW-A model for $T_{\text{MMCW-A}}$ time with
  initial solution $(\hat{x}, \hat{v}, (\hat{f}, \hat{y}))$ and objective value $val$ as input;
Set $(x', v', (f', y')) \leftarrow$ output solution;
Set $val \leftarrow$ output objective value;
Set $z'_{lmf} \leftarrow \mathbb{1}_{\{f'_{lm} = f\}} \quad \forall\, l \in \mathcal{L}, \forall\, m \in \mathcal{M}_l, \forall\, f \in \mathcal{F}_{lm}$;
Run single-phase IPBLS (Algorithm 3) using MMCW model for $T_{\text{MMCW}}$ time with initial
  solution $(x', v', z')$ and objective value $val$ as input;
Set $(\hat{x}, \hat{v}, \hat{z}) \leftarrow$ output solution;
Set $val \leftarrow$ output objective value;
return $(\hat{x}, \hat{v}, \hat{z})$, $val$
```

demonstrate the impact of the conservatism hyperparameters on the on-time probabilities of the consolidation plans generated by the MMCW and MMCW-A models; and (iii) assess the value of using the single- and two-phase IPBLS heuristic approaches to solve large-scale, realistic instances of these models.

The load plan models and IPBLS heuristic approaches were coded in Python 3.7 using Gurobi 9.1.1 as the MIP solver. We set the Gurobi MIPFocus parameter to focus on finding feasible solutions when solving the restricted MIPs within Algorithm 3 and used the default setting (balancing feasibility and optimality) when solving the complete MIPs. All experiments were run on a Linux computing cluster, which uses HTCondor 8.8.12 for job management. Each node in the cluster uses multi-core 2.4 GHz processors with 8 GB of RAM each.

The IPBLS heuristics used to solve model instances were tuned using experiments, and we now provide some details. When selecting the subsets of routes $\mathcal{R}^{(t)}$ to free for optimization at an iteration of the single- and two-phase IPBLS heuristics, we set $\alpha = 0.3$ to balance the MIP solution work required per iteration with the number of iterations. Other values for α (i.e., $\alpha = 0.2, 0.25, 0.35, 0.4$) were tested and found to either not make a significant improvement or reduced the heuristic performance by either restricting too few or too many route variables. It may be useful in practice to have α vary as Algorithm 3 proceeds; larger values of α can be used in later iterations to intensify the search. We additionally set a solve time limit of 5 minutes per MIP to ensure a large number of iterations within the total heuristic time limit of 12 hours while also providing sufficient solve time for the restricted MIP such that the final restricted MIP gap was less than 5%. The number of non-improving neighborhood searches allowed before switching neighborhood selection methods is set to $N = 5$. We experimented with slightly different values (e.g., $N = 3$, 4, 6, and 7) but did not observe a significant effect on the results. The exact value of this parameter is not necessarily important; the purpose of alternating selection methods is to help the search escape local minima. The value of N should be small enough such that it provides the heuristic with enough time to use both randomized neighborhoods multiple times, but also large enough to allow a thorough search of the neighborhoods before switching to the next.

5.1. Middle-Mile Network Instances

The instances used in this study are synthetic but have been derived from historical demand data provided by a large U.S.-based e-commerce retailer that partnered with our research team. Each instance uses a planning horizon of one week. Shipment demand originates from locations in a set $\mathcal{N}_O$ of vendors (VND) and FCs. Shipment destinations are locations in a set $\mathcal{N}_D$ of LMD facilities. We categorize the vendors and LMD facilities into three size groups depending on the amount of volume these locations send or receive, respectively. The distributions of vendors and LMD facilities across the size groups are representative of those in our partner's network. The set $\mathcal{N}_H$ of facilities used for intermediate shipment transfer in these instances is limited to the FCs; each instance has 8 such facilities. We create 9 groups of instances of increasing size that differ in the number of vendors, LMD facilities, and demand commodities. Within each group, we build 5 instances with different VND and LMD locations and commodity sets.

Attributes of the instances are summarized in Table 1. The table includes the number of VND and LMD types, as well as the average number of truckloads (TL) originating at VNDs and FCs and destined for LMDs, the number of commodities, and the number of lanes and routes for each instance of each group. See Appendix B for additional details on the instances, including characteristics of the average flow between location types. Figure 5 shows the vendor, FC, and LMD locations for Group 4 - Instance 1; FC locations are identical in all groups and instances.

Direct freight transportation legs exist from each vendor location to each LMD facility that receives shipments from the vendor. Furthermore, a leg exists between each vendor and each FC, from FC to FC, and from each FC to each LMD facility. The truckload freight mode is available for all these legs. LTL freight (and weight bucket modes) is allowed only on direct legs and FC-to-LMD facility legs, since these restrictions most closely resemble the operations of our e-commerce partner. For the truckload mode, trailer capacity is set at 12,000 pounds since load cube is typically the binding size constraint for e-commerce shipments. LTL transportation is modeled with three

Table 1 **Instance characteristics.**

Gr.	# of Locations / Average Volume in # of Truckloads (TL)							Comm.	Inst.	Lanes	Routes
	Small VND	Medium VND	Large VND	FC	Small LMD	Medium LMD	Large LMD	$\vert\mathcal{K}\vert$		$\sum_{l\in\mathcal{L}}\vert\mathcal{M}_l\vert$	$\vert\mathcal{R}\vert$
1	10 0.1 TL	2 0.5 TL	0 0 TL	8 2.5 TL	5 3.1 TL	2 3.5 TL	0 0 TL	106	1	548	506
									2	524	506
									3	547	494
									4	555	492
									5	551	500
2	15 0.3 TL	5 1.4 TL	2 3.1 TL	8 6.6 TL	10 3.8 TL	5 4.4 TL	2 5.1 TL	371	1	1,706	1,725
									2	1,704	1,713
									3	1,705	1,733
									4	1,722	1,713
									5	1,717	1,749
3	25 0.6 TL	10 2.4 TL	5 5.4 TL	8 12.1 TL	20 4.2 TL	10 4.9 TL	5 5.8 TL	1,176	1	5,127	5,528
									2	5,131	5,478
									3	5,118	5,512
									4	5,133	5,550
									5	5,132	5,556
4	50 0.9 TL	20 3.9 TL	10 9.0 TL	8 19.3 TL	30 5.6 TL	15 7.1 TL	10 9.4 TL	3,278	1	13,894	15,342
									2	13,886	15,374
									3	13,884	15,344
									4	13,890	15,368
									5	13,887	15,362
5	75 1.1 TL	30 4.6 TL	15 10.5 TL	8 22.8 TL	40 6.7 TL	20 8.7 TL	10 12.0 TL	5,992	1	25,083	27,968
									2	25,096	28,092
									3	25,093	28,032
									4	25,095	28,176
									5	25,094	28,016
6	100 1.4 TL	40 5.9 TL	20 13.4 TL	8 28.8 TL	50 7.9 TL	25 106 TL	15 14.4 TL	10,044	1	41,655	47,092
									2	41,656	47,180
									3	41,655	47,032
									4	41,651	46,934
									5	41,656	47,216
7	150 1.5 TL	50 6.2 TL	25 14.0 TL	8 30.8 TL	60 8.7 TL	30 11.9 TL	15 16.5 TL	16,023	1	66,116	75,153
									2	66,116	75,177
									3	66,115	74,927
									4	66,115	75,033
									5	66,115	75,155
8	250 1.8 TL	75 7.3 TL	40 16.6 TL	8 35.9 TL	70 12.2 TL	35 17.0 TL	20 24.3 TL	30,263	1	124,228	141,947
									2	124,228	142,009
									3	124,228	141,955
									4	124,228	141,837
									5	124,227	142,029
9	375 1.9 TL	100 7.9 TL	50 18.0 TL	8 39.8 TL	80 15.5 TL	40 21.8 TL	20 31.0 TL	47,964	1	196,336	225,210
									2	196,335	225,014
									3	196,335	225,306
									4	196,336	225,216
									5	196,336	225,318

weight buckets with minimum capacities of 0, 2,000, and 2,700 pounds and maximum capacities of 2,000, 2,700, and 4,000 pounds, respectively. The cost of a truckload was derived from historical freight costs provided by our partner. The LTL cost buckets were then derived with the intuition that moving more than a third of a truckload using LTL is typically more expensive than moving

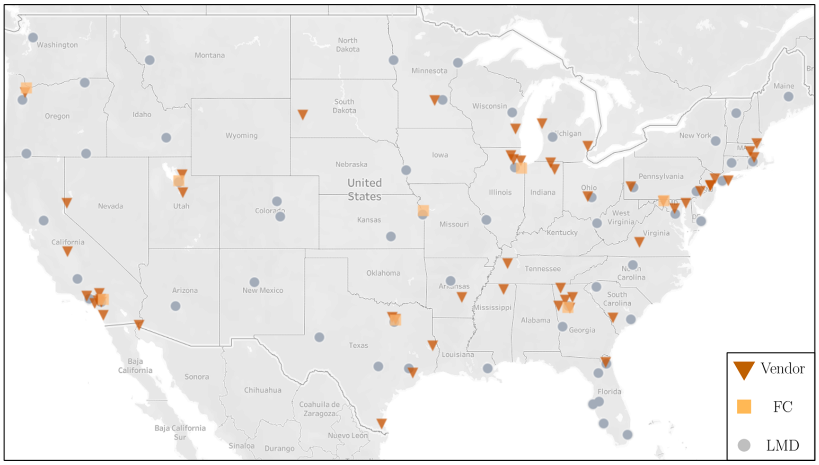


Figure 5 **Locations for Group 4 - Instance 1.**

that load by truckload. A summary of freight costs is provided in Table 2, where d represents the distance of the leg; see Appendix B for an illustration of the freight costs of a 500-mile leg.

Table 2 **Freight mode costs for a single load.**

Mode	Weight (lbs)	Cost
TL	$0 < w \leq 12{,}000$	$750 + 1.27d$
LTL$_1$	$0 < w \leq \ \ 2{,}000$	$0.05(750 + 1.27d) + w(0.234 + 0.0004d)$
LTL$_2$	$2{,}000 < w \leq \ \ 2{,}700$	$0.05(750 + 1.27d) + 2000(0.234 + 0.0004d)$
LTL$_3$	$2{,}700 < w \leq \ \ 4{,}000$	$0.8w(0.234 + 0.0004d)$

We allow up to 40 truckloads on each leg during the week. On the other hand, we limit LTL shipping to 5 loads per week, which represents sending a single load per weekday; once more capacity is needed, truckloads will be required.

For each instance, we generate sets $\mathcal{R}_k$ with the 5 most operationally-reasonable route options for each commodity k using guidelines representing a more flexible version of the methods currently used by our industry partner. The direct, single-leg route connecting the commodity origin to destination is always included; selecting this route represents the decision to exclude this commodity from any consolidated middle-mile loads and to simply send these shipments direct. The remaining 4 possible route options are: (i) the shortest distance two-leg route using a single transfer FC, (ii) the two-leg route transferring at the FC closest to the vendor, (iii) the two-leg route transferring at the FC closest to the LMD facility, and (iv) a three-leg route transferring at the FCs in (ii) and (iii), if they are not identical. If some routes are identical, duplicates are removed, resulting in commodities with fewer than 5 route options. Note that the number of routes is linear in the number of commodities; thus, the computational burden to generate such route sets is minimal. The restriction to use at most two transfer locations is common in practice for large middle-mile operations; this was the case for the large e-retailer we partnered with.

Without loss of generality, we assume that each unique commodity k represents shipment volume from origin o_k to destination d_k with identical lead-time targets; if shipments between these facilities have potentially different lead-time targets, additional commodities could be defined. Since we are unable to share actual target delivery lead times from our e-commerce partner, we generate realistic substitutes by randomly perturbing promised lead times between various geographic origin-destination pairs with a multiplicative factor drawn uniformly from $[0.8, 1.2]$. These lead-time targets are then used to calculate the maximum allowable waiting delay $\hat{W}_r$ for each route by subtracting its fixed transit time and FC transfer processing time(s), when applicable. Time-infeasible routes (i.e., those that can never meet their lead-time requirement) are removed prior to solving all optimization models. Allowable waiting delay constraints are generated with a conservatism level of $\rho_r = 0.5$ (to meet lead-time requirements in expectation) in the computational experiments to follow, except those in Section 5.4 where the results under various on-time probabilities are compared.

To reduce the computational burden when solving the models, the freight modes, load dispatch frequencies, and related costs for all direct routes were determined in a pre-processing step. Importantly, this allows the cost of assigning a commodity to its direct route r' to be included entirely in the route cost coefficient $C_{r'}$; direct route legs are thus excluded from the set $\mathcal{L}$, substantially reducing the number of decision variables and related constraints.

5.2. Exact MIP versus Heuristic Solution Approach Performance

First, we present results that verify that our single- and two-phase IP-based heuristic solution approaches are effective at solving realistically-sized problems. To evaluate MIP gaps associated with solutions generated by the different approaches, we compute a best-known lower bound by allowing the full MIP model to run for 2 weeks with the Gurobi MIPFocus parameter set to focus on improving the lower bound. Table 3 shows the average performance across instances for each group, comparing the average solution objective function values resulting when solving the full MIP directly versus when solving using the single-phase IPBLS heuristic; the far right column reports the percentage improvement in objective value when using the heuristic. The time until the best objective found by the heuristic is also reported as Time to Obj (hr). This metric highlights both that the heuristic can work well for smaller instances with a shorter run time and how the required run time quickly increases as the instance size increases.

Although the full MIP can optimally solve Group 1 instances, the quality of the solutions produced by the full MIP unsurprisingly degrades as the instance size increases, where both the MMCW-A and MMCW solutions have MIP optimality gaps greater than 50% for the largest instances. On the other hand, we see that the heuristic approach produces high-quality solutions for

Table 3 Comparing MIP vs single-phase IPBLS heuristic performance.

Group	2-week MIP	MIP		IPBLS Heuristic			Heuristic
	Lower Bound	12-hr Obj	MIP Gap	12-hr Obj	MIP Gap	Time to Obj (hr)	Improvement
1	$120,000	$120,000	0.0%	$120,000	0.0%	0.1	0.0%
2	$353,000	$359,000	1.7%	$362,000	2.5%	0.9	-0.8%
3	$717,000	$767,000	6.5%	$778,000	7.8%	7.1	-1.4%
4	$1,541,000	$1,704,000	9.6%	$1,712,000	10.0%	12.0	-0.5%
5	$2,275,000	$2,653,000	14.2%	$2,587,000	12.0%	12.0	2.5%
6	$3,702,000	$4,774,000	22.4%	$4,159,000	11.0%	12.0	12.9%
7	$4,990,000	$6,908,000	27.8%	$5,639,000	11.5%	12.0	18.4%
8	$12,457,000	$15,543,000	28.6%	$12,457,000	10.9%	12.0	19.9%
9	$20,788,000	$45,657,000	54.5%	$22,911,000	9.3%	12.0	49.8%

(a) MMCW-A

Group	2-week MIP	MIP		IPBLS Heuristic			Heuristic
	Lower Bound	12-hr Obj	MIP Gap	12-hr Obj	MIP Gap	Time to Obj (hr)	Improvement
1	$118,000	$118,000	0.0%	$118,000	0.0%	0.1	0.0%
2	$299,000	$311,000	3.9%	$311,000	4.0%	3.3	-0.1%
3	$625,000	$686,000	8.9%	$693,000	9.8%	9.5	-1.0%
4	$1,355,000	$1,536,000	11.7%	$1,522,000	11.0%	12.0	0.9%
5	$2,046,000	$2,359,000	13.2%	$2,323,000	11.9%	12.0	1.5%
6	$3,267,000	$4,226,000	22.7%	$3,780,000	13.6%	12.0	10.6%
7	$4,331,000	$6,046,000	28.4%	$5,096,000	15.0%	12.0	15.7%
8	$9,171,000	$19,633,000	53.3%	$12,512,000	26.7%	12.0	36.3%
9	$15,327,000	$39,720,000	61.4%	$25,969,000	41.0%	12.0	34.6%

(b) MMCW

all instance sizes within the allowed time limit, especially when using the MMCW-A models. Given that the MMCW-A model is a restriction of the MMCW model, the similar (or better) MMCW-A objective and solution quality for larger instances demonstrates why it is effective to solve the MMCW-A model to create good warm-start solutions for the two-phase solution approach. We present results to confirm this idea in Table 4. We additionally give example plots in Figure 6 from both Group 8 and 9 instances that visually demonstrate the effectiveness of the two-phase approach. All other Group 8 and 9 instance plots exhibit nearly identical behavior.

Table 4 MMCW-A and MMCW solved using single-phase IPBLS compared to the two-phase IPBLS approach for solving MMCW.

Group	MMCW MIP 2-week LB	MMCW-A		MMCW		8-hr MMCW-A + 4-hr MMCW	
		IPBLS 12-hr Obj	MIP Gap	IPBLS 12-hr Obj	MIP Gap	IPBLS 12-hr Obj	MIP Gap
8	$9,171,000	$12,457,000	26.4%	$12,512,000	26.7%	$10,938,000	16.1%
9	$15,327,000	$22,911,000	33.1%	$25,969,000	41.0%	$19,730,000	22.3%

We found that a good distribution of solve time limits for the two-phase IPBLS approach is to allocate 8 hours to the solution of the MMCW-A model and then 4 hours to the solution of the MMCW model; this allows the MMCW-A objective value to reach a plateau where fewer improvements can be easily identified, while still providing sufficient solve time to allow the MMCW model to find improvements. Using these parameters, the MMCW solution objective improves by 13% for Group 8 and 24% for Group 9, while the MIP gaps improve by 39% and 45%, respectively.

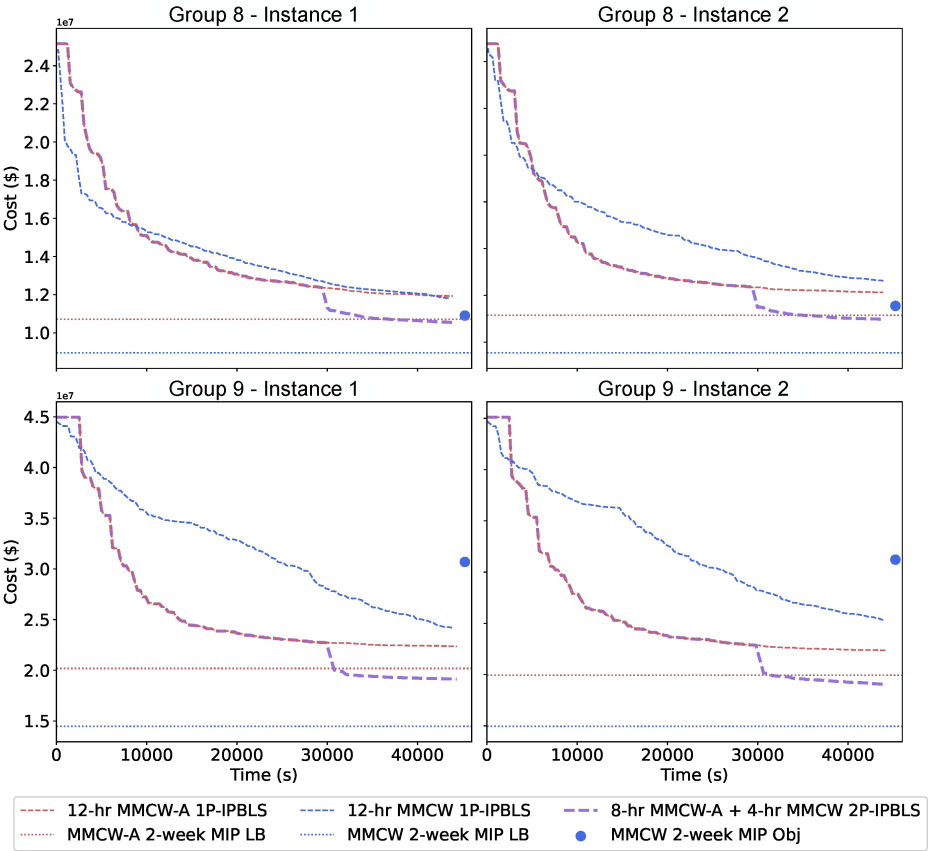


Figure 6 **Example plots for our single- and two-phase IPBLS solution approaches for Groups 8 and 9.**

We make two observations when analyzing the example plots in Figure 6: (i) the drop in objective value when switching from the MMCW-A model to the MMCW model after 8 hours, representing the improvement in objective function when restrictions on leg load dispatch frequencies are relaxed (i.e., individual leg waiting delays need not be less than the equally-distributed allowable waiting delay $2\frac{\hat{W}_r}{|r|}$); and (ii) the 8-hr MMCW-A + 4-hr MMCW solution objective value drops below the 2-week MIP lower bound for the MMCW-A model, demonstrating that utilizing the two-phase approach allows us to obtain a better solution than we could have obtained (at optimality) when solving the MMCW-A model alone.

5.3. The Effect of Constraining Lead Time

We now study the effect of adding lead-time constraints when optimizing a middle-mile network design using the MMCW model. We present in Table 5 results from solving the MMC model and the MMCW model with the 12-hr single-phase IPBLS heuristic approach on the same instances. To obtain better solutions for the larger instances in groups 8 and 9, we use the 12-hr two-phase IPBLS heuristic to solve the MMCW model. All row values represent averages across the 5 instances within each group and we abbreviate Volume-Weighted as Vol-Wtd and Average Load Dispatch Frequency as Avg Load Disp Freq in the column titles.

Table 5 Comparison of the MMC and MMCW model solutions.

	IPBLS 12-hr Objective	Vol-Wtd Route Length	Vol-Wtd Route Length Variance	Avg Load Disp Freq (#/week)		Loads/Week		Vol-Wtd Utilization
				LTL	TL	LTL	TL	TL
Group 1								
MMC	$90,000	1.5	0.3	1.0	1.1	30	45	80%
MMCW	$118,000	1.5	0.3	1.3	1.9	70	40	67%
Group 2								
MMC	$256,000	1.6	0.4	1.0	1.2	20	140	83%
MMCW	$311,000	1.7	0.4	1.1	1.8	70	170	77%
Group 3								
MMC	$573,000	1.7	0.4	1.0	1.3	30	310	88%
MMCW	$693,000	1.8	0.4	1.2	2.0	140	370	82%
Group 4								
MMC	$1,249,000	1.7	0.3	1.0	1.4	60	680	91%
MMCW	$1,522,000	2.0	0.4	1.4	2.4	230	843	85%
Group 5								
MMC	$1,888,000	1.7	0.3	1.0	1.5	120	1,060	91%
MMCW	$2,323,000	2.1	0.4	1.6	3.3	300	1,360	86%
Group 6								
MMC	$2,956,000	1.8	0.2	1.0	1.7	60	1,300	92%
MMCW	$3,780,000	2.1	0.4	2.0	3.8	540	2,180	84%
Group 7								
MMC	$3,862,000	1.8	0.2	1.0	1.7	60	2,230	93%
MMCW	$5,096,000	2.0	0.4	2.0	3.4	1,320	2,830	81%
Group 8								
MMC	$6,575,000	1.9	0.1	1.0	2.1	50	3,860	94%
MMCW-A+MMCW	$10,938,000	2.1	0.3	2.5	6.3	2,000	6,020	61%
Group 9								
MMC	$9,230,000	1.9	0.1	1.0	2.3	100	5,490	93%
MMCW-A+MMCW	$19,730,000	2.2	0.3	2.5	7.7	2,610	10,310	41%

As expected, we observe, by comparing the IPBLS 12-hr objectives, that the best solutions to the MMCW models require more total cost than solutions to the MMC models, and thus total middle-mile cost increases sometimes significantly once lead-time constraints are added. The volume-weighted route lengths show that the MMC solution routes most commodities through one FC and then on to the destination. Load dispatch frequencies provide only enough capacity for the shipment volumes on each leg, leading to dispatch rates of one or two loads per week on most legs. This results in high load volume-weighted utilization and low costs per ton-mile. However, when lead-time constraints are enforced, the design must better utilize consolidation lanes to achieve both cost scale economies and to meet lead-time constraints. The result is an increase in the average load dispatch frequencies; the table shows increases, sometimes dramatic, in both LTL and truckload lane dispatch frequencies averaged across all lanes with positive frequency. We also observe an increase in the use of LTL loads to move shipments, clearly shown by the increase in the absolute number of loads dispatched per week; note even for the smallest instances, we see at least a doubling, often tripling, in the number of LTL loads per week. This occurs largely because some commodities cannot find a good consolidation path (even with increased truckload frequencies) that meets their lead-time constraints; these commodities must be served with frequent LTL shipments on direct legs.

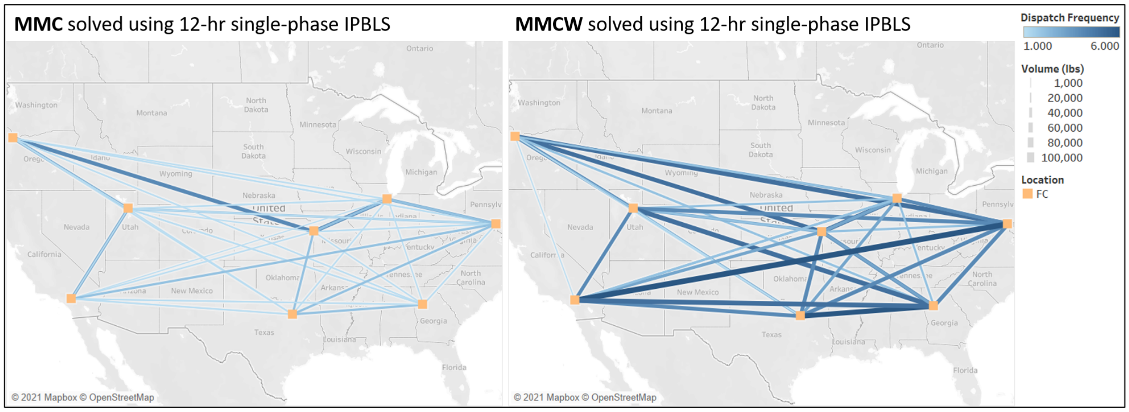

MMC solved using 12-hr single-phase IPBLS
MMCW solved using 12-hr single-phase IPBLS
Dispatch Frequency
1.000
6.000
Volume (lbs)
1,000
20,000
40,000
60,000
80,000
100,000
Location
FC
© 2021 Mapbox © OpenStreetMap
© 2021 Mapbox © OpenStreetMap

Figure 7 **Comparison of FC-to-FC truckload lane volume and load dispatch frequencies for the MMC and MMCW models (Group 4 - Instance 1).**

It is interesting to note that when lead-time constraints are enforced, we observe an increase in the volume-weighted average route length, measured in number of middle-mile legs per shipment volume. This can be explained by the solution aiming to mitigate the cost increase from setting higher load frequencies by consolidating more commodities. Indeed, by adding lead-time constraints, load frequencies must increase to reduce waiting times between dispatches on some lanes. However, increasing frequencies on lanes with low volume is significantly more expensive compared to lanes with higher volume. As a result, the solution assigns more volume to two-leg routes and three-leg routes that include FC-to-FC truckload lanes. Interestingly, high dispatch frequencies on consolidation lanes can reduce waiting delay enough to offset the higher transit times that result when shipments follow longer transit-time geographic paths. In Figure 7, we observe this increase in volume and load dispatch frequency on the FC-to-FC truckload lanes. In the figure, lanes are represented by blue lines, where a thicker line indicates more volume (across all commodities) and a darker shade of blue indicates a higher load dispatch frequency. Surprisingly, dispatch frequencies on consolidation lanes are increased so significantly that many of the commodities that use routes with more legs meet their lead-time requirements with more slack time. Figure 8 shows the distributions of the allowable waiting delay slack (i.e., $\hat{W}_r$ net the expected waiting delay given the solution) for each route type when adding MMCW lead-time constraints compared to the MMC model. Interestingly, allowable waiting delay slack increases on average as the route length increases, demonstrating the powerful reductions in waiting time possible when moving large shipment volumes on consolidation lanes.

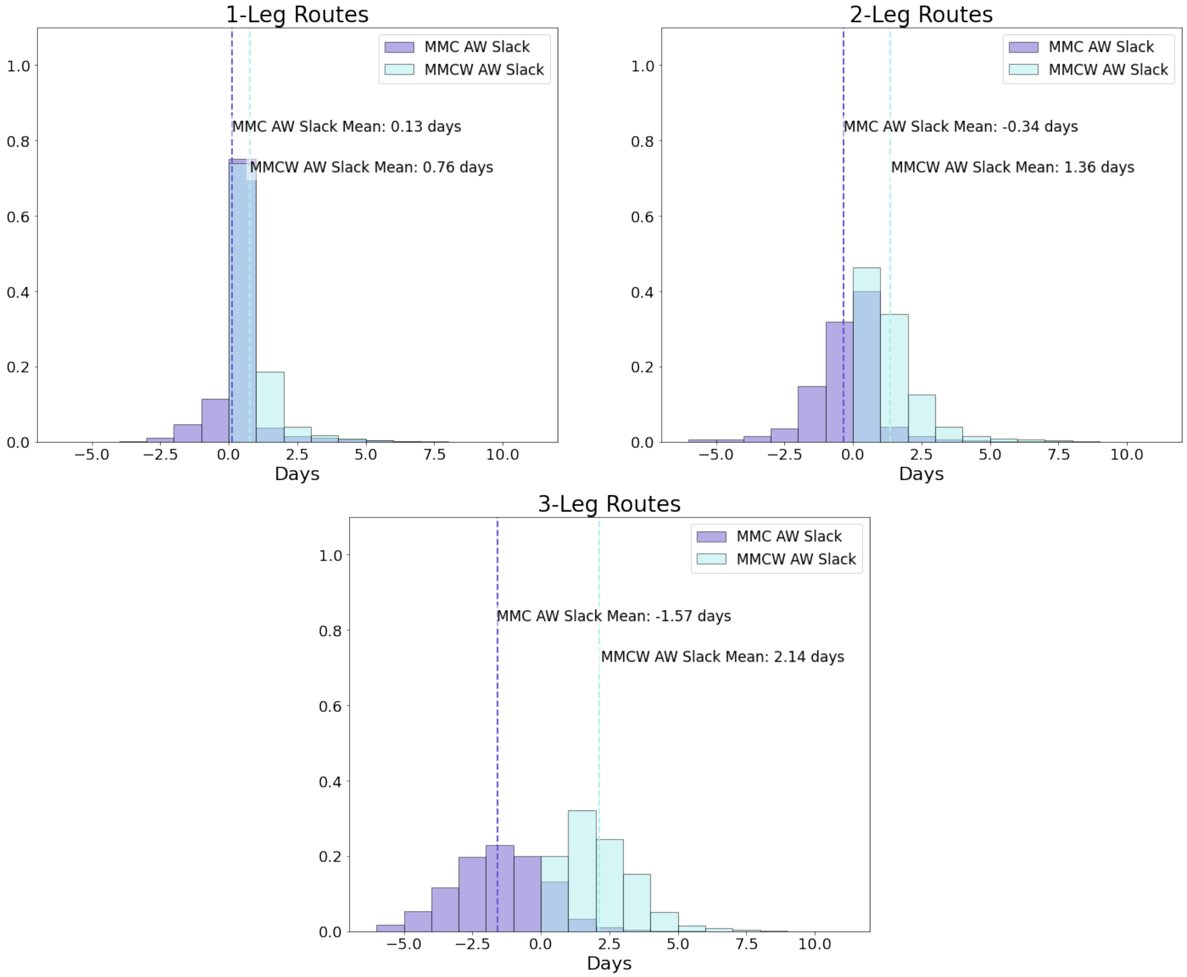


Figure 8 Distributions of MMC and MMCW allowable waiting (AW) delay slack for the selected routes across all instances categorized by route type.

Finally, we show two examples from the Group 4 instances in Figure 9 of a particular commodity being re-routed to increase consolidation between FCs. The arrows in the figure represent the route to which this commodity was assigned in that solution. Commodity (a) was assigned to the direct route in the MMC solution, but instead consolidates at the nearest FC along with other commodities in the MMCW solution. Commodity (b) switches from a two-leg route to a three-leg route; the second leg in the MMC solution is no longer used in the MMCW solution, as the model is able to reduce costs by consolidating all commodities to use the FC-to-FC transfer. Although the MMCW model selects a longer route (with additional transit time) for commodities (a) and (b), the increased load dispatch frequencies result in a much smaller waiting delay, which in turn reduced the total lead time.

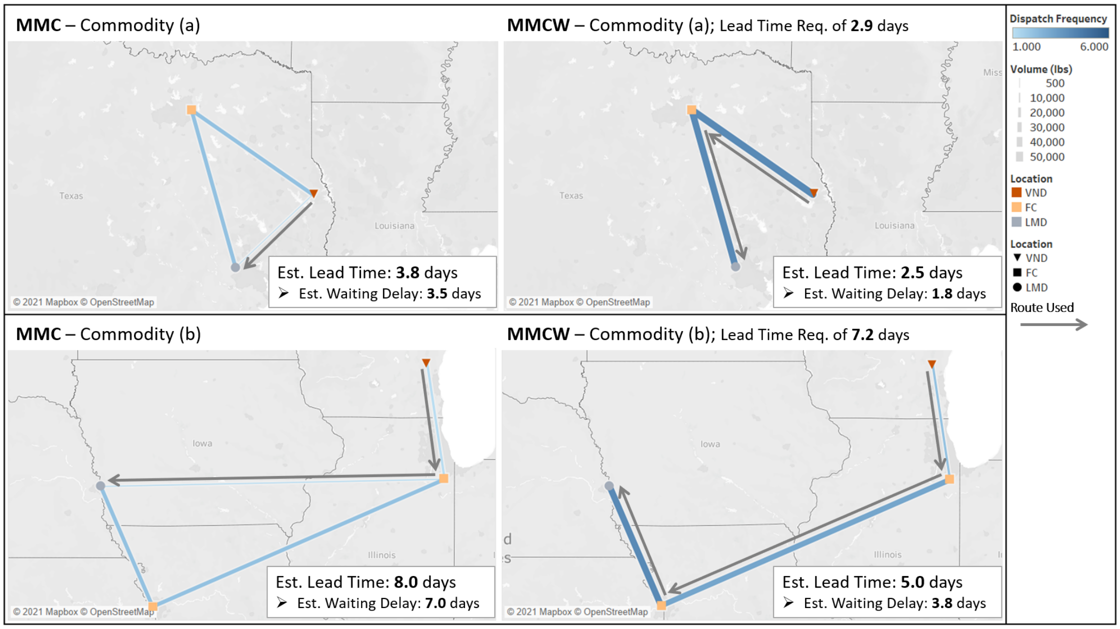


Figure 9 **Examples of commodities opting for longer routes in MMCW (Group 4 - Instance 1).**

Overall, we find that solutions of the MMCW model utilize more consolidation lanes to offset the increased costs associated with higher load dispatch frequencies. Although this leads to longer routes (both in number of legs and miles) on average, there is still a significant decrease in expected lead times for commodities.

5.4. Analysis of Conservatism

Next, we analyze the effect of conservatism on the solution, specifically looking at the trade-off between cost and on-time probability. As discussed in Section 3.3, a minimum on-time probability p can be specified to set conservatism levels ρ_r either for each individual route (for MMCW models) or by route type (for MMCW-A models). Of course, the network designs generated by these models will result in higher service levels than the minimum on-time probability specified. One way we choose to measure service level in a solution is by calculating the volume-weighted expected on-time probability (vOTP) of a solution as follows:

$$\text{vOTP} = \frac{\sum_{k\in\mathcal{K}} \mathbb{P}\left(\sum_{l\in r} W_l \leq \hat{W}_r\right) V_k}{\sum_{k\in\mathcal{K}} V_k},$$

where the on-time probability of an individual commodity $\mathbb{P}\left(\sum_{l\in r} W_l \leq \hat{W}_r\right)$ when using route r, is calculated using (4), the assigned load dispatch frequencies from the solution, and the allowable waiting delay $\hat{W}_r$. As another service level performance metric, we calculate the volume-weighted maximum lateness of a solution. To calculate this worst-case metric, every commodity is assumed to experience waiting time of a full headway on each leg they traverse. The difference between the sum of the maximum possible waiting delays across the route and the commodity's allowable waiting delay is the maximum lateness the commodity may experience. If the difference is negative, the commodity will never be late and the maximum lateness is zero. These maximum lateness values and the commodity volumes are used to calculate the volume-weighted maximum lateness of a solution.

The results for two mid-sized groups with four different values of on-time probability guarantees (i.e., Min p) are shown in Table 6, as well as results for the associated MMC solution. Note that in this section, MMCW-A model results are provided primarily for comparative analysis; in general, this stand-alone model produces results that are too costly and conservative to be useful in practice. Each row represents the average measure for the 5 instances within each group when solving the models using the 12-hr single-phase IPBLS approach. We again abbreviate Volume-Weighted as Vol-Wtd and Average Load Dispatch Frequency as Avg Load Disp Freq in the column titles.

The lead-time constraints in the MMCW-A model, compared to those in the MMCW model, are stricter for two- and three-leg routes, which leads to higher costs and more conservative average load dispatch frequencies. Naturally, this also leads to higher expected on-time probabilities, as measured by the increase in vOTP, and reduced maximum lateness values in solutions found using the MMCW-A model. A retail shipper is likely to evaluate a consolidation network design using two main metrics: (i) cost and (ii) service level. In Table 6, we first observe that design solutions resulting from only minimizing cost would result in a vOTP of 47% and a maximum lateness of about 4 days. We then see a significant improvement in both vOTP and maximum lateness when adding lead-time constraints, even for the case where all commodities are only guaranteed to be on time at least 50% of the time. While the lead-time constraints (7) for MMCW (resp. (10) for MMCW-A) guarantee an on-time probability p for all commodities, we find that in nearly all cases, the resulting vOTP exceeds p. This excess is due to some load frequencies being driven by high volume commodities through constraints (1d) or by commodities with tight deadlines requiring more frequent dispatches. When balancing cost and service level, using the MMCW model with $p = 0.7$ appears to be a reasonable choice. For both groups, commodities have an expected on-time probability of 96% and maximum lateness of less than 0.25 days for only a roughly 10% increase in cost compared to the case when $p = 0.5$.

Table 6 **Comparing different service levels for both MMCW and MMCW-A solved using 12-hr IPBLS.**

Min p	Model	12-hr IPBLS Obj	vOTP	Vol-Wtd Max Lateness (days)	Vol-Wtd Route Length	Avg Load Disp Freq (#/week)		Loads/Week		Vol-Wtd Util
						LTL	TL	LTL	TL	TL
Gr 5										
0	MMC	\$1,888,000	0.47	4.25	1.7	1.0	1.5	120	1,060	91%
0.5	MMCW	\$2,323,000	0.87	0.83	2.1	1.6	3.3	300	1,360	86%
	MMCW-A	\$2,587,000	0.93	0.47	2.1	1.9	3.9	700	1,430	82%
0.6	MMCW	\$2,436,000	0.92	0.47	2.1	2.1	3.7	430	1,410	83%
	MMCW-A	\$2,760,000	0.97	0.20	2.1	2.4	4.4	850	1,490	79%
0.7	MMCW	\$2,523,000	0.96	0.24	2.2	2.3	4.7	490	1,510	86%
	MMCW-A	\$2,842,000	0.98	0.11	2.2	2.5	5.2	870	1,570	79%
0.8	MMCW	\$2,662,000	0.98	0.12	2.3	2.6	5.3	620	1,580	84%
	MMCW-A	\$2,943,000	0.99	0.05	2.3	2.6	6.2	960	1,680	80%
0.9	MMCW	\$2,844,000	1.00	0.05	2.3	2.6	5.8	720	1,680	78%
	MMCW-A	\$3,103,000	1.00	0.02	2.4	3.0	7.1	870	1,840	76%
1.0	MMCW	\$3,207,000	1.00	0.00	2.3	3.0	7.3	870	1,930	72%
	MMCW-A	\$3,753,000	1.00	0.00	2.3	3.3	8.9	1,290	2,150	64%
Gr 6										
0	MMC	\$2,956,000	0.47	3.93	1.8	1.0	1.7	60	1,300	92%
0.5	MMCW	\$3,780,000	0.87	0.74	2.1	2.0	3.8	540	2,180	84%
	MMCW-A	\$4,159,000	0.95	0.26	2.2	2.5	5.5	1,030	2,390	86%
0.6	MMCW	\$3,979,000	0.93	0.38	2.1	2.6	4.6	750	2,310	82%
	MMCW-A	\$4,429,000	0.98	0.12	2.3	2.6	6.4	1,250	2,540	82%
0.7	MMCW	\$4,259,000	0.96	0.21	2.2	2.7	5.4	1,020	2,460	80%
	MMCW-A	\$4,632,000	0.99	0.06	2.4	3.0	7.4	1,270	2,730	80%
0.8	MMCW	\$4,552,000	0.98	0.10	2.3	2.9	6.4	1,130	2,650	76%
	MMCW-A	\$4,930,000	0.99	0.02	2.4	3.1	8.5	1,330	2,970	75%
0.9	MMCW	\$4,783,000	1.00	0.04	2.4	3.1	7.6	1,030	2,910	74%
	MMCW-A	\$5,303,000	1.00	0.01	2.4	3.5	9.8	1,170	3,310	69%
1.0	MMCW	\$5,560,000	1.00	0.00	2.4	3.4	9.4	1,430	3,370	64%
	MMCW-A	\$6,526,000	1.00	0.00	2.4	3.4	11.9	1,890	4,030	57%

6. Conclusion and Future Work

In this article, we studied a middle-mile consolidation network design problem to improve the service level and outbound logistics cost of large e-commerce retailers. Specifically, we considered the problem of capacity planning for moving customer shipments from fixed stocking locations to LMD partners at minimum cost while satisfying customer promised delivery times. We proposed three MIPs where both input demands and planned load decisions are expressed as constant rates per time, extending traditional flat network SND models. First, the MMC base model handles cases where shipment lead times can be completely determined by the legs and transfer terminals within each route. To better account for the shipment waiting delays incurred between load dispatches, we introduced chance constraints that guarantee lead-time requirements are met with a desired probability specified by the shipper. We approximated these chance constraints using hyperparameterized nonlinear constraints, which we reformulated as linear constraints using binary variables.

The second MIP, the MMCW model, was obtained by adding these new constraints to the MMC model. Third, we developed a simpler restricted MIP, the MMCW-A model, that individually constrains leg waiting delays to satisfy lead-time requirements with the desired probability.

To find high-quality solutions to these large-scale MIPs, we developed an effective single-phase IPBLS heuristic that iteratively improves an incumbent solution by optimizing over a smartly selected subset of commodities using two neighborhood selection methods. For the largest problem instances, we also proposed a two-phase IPBLS heuristic that first runs the single-phase IPBLS on the MMCW-A model, and then further improves the incumbent solution using the single-phase IPBLS on the MMCW model.

We then conducted an extensive computational study using data from a large U.S.-based e-commerce partner to demonstrate the impact of tight lead-time constraints on the structure of the consolidation network designs and their concomitant operating costs. Notably, we observed that tighter and more conservative lead-time constraints lead to solutions with increased shipment consolidation and higher dispatch frequencies on selected key transportation lanes. Such solutions trade off higher shipment transit times with significantly reduced shipment waiting times to meet lead-time constraints at lower cost. Finally, we found that the single- and double-phase IPBLS heuristics provide a significant improvement over the solutions obtained directly from optimization solvers, especially for large real-world problem instances.

Although we believe that the solutions produced by our approach are of high quality, it is difficult to compute tight objective lower bounds for larger problem instances. Future work could focus on improving these lower bounds. Furthermore, the modeling framework in this paper is largely deterministic since input demand rates are considered to be known and time-homogeneous and planned dispatches are assumed to occur with constant known headways. Robust extensions of this work that relax these assumptions could be useful in practice.

A natural extension to this work is to incorporate transportation equipment management. In our current work, the shipper determines how to economically consolidate loads and meet shipment time requirements when outsourcing loads to third-party carriers operating their own equipment. However, the shipper may decide to acquire a dedicated fleet of trailers to ensure capacity availability and potentially further reduce costs. In this case, the optimization model would require equipment balance constraints, which equalize the total inbound and outbound load frequencies for the truckload mode at each location. If the dedicated fleet were limited in size, it would also likely be necessary to include an approach for modeling total truckload trailer availability (and possibly also to distinguish between both a dedicated truckload mode and a third-party outsourced truckload mode). The addition of these constraints would likely lead to a more difficult-to-solve model; thus, research is necessary to determine if an appropriate local search procedure could be developed to find high-quality solutions in this case.

Another natural extension of this work is to seek methods for determining a detailed timed schedule of load dispatches for a planning horizon. After solving our model, the tactical consolidation plan given by the set of selected routes for all commodities can be fixed as an input to a detailed scheduling approach that uses a time-expanded network model to determine dispatch dates and times for a set of loads. Such an approach would require more precise forecasts of commodity demand at specific times during the planning horizon and could be used to more accurately determine the number of loads required to transfer all demand between origins and destinations to meet lead-time requirements. One could also consider the problem with flexible origins and destinations, where the shipper can decide the origin of shipments containing items held in stock at multiple locations and can also select an LMD destination from potentially multiple locations with different cost and lead-time implications. For example, dropping a shipment at the local terminal of an LMD partner might be cheaper than using the middle-mile network for some shipments; forcing all commodities to find an effective consolidation path that meets lead-time constraints through the middle-mile network may be overly restrictive. Additionally, our models can be used to reallocate items in stock among the FCs by leveraging the unused truckload capacity in the selected consolidation plan in order to reduce future lead times.

Appendix A: Derivation of (11)

We provide the complete derivation of (11) below. For every route $r \in \mathcal{R}$ and every set of frequencies $f \in \mathbb{Z}_{>0}$ satisfying $\frac{1}{f_l} \leq \frac{\hat{W}_r}{\rho_r|r|}$ for every $l \in r$, we have:

$$\begin{aligned}
\mathbb{P}\left(\sum_{l\in r} W_l \leq \hat{W}_r\right) &\geq \frac{1}{|r|!\prod_{l\in r}\frac{\hat{W}_r}{|r|\rho_r}}\sum_{J\subseteq r}(-1)^{|J|}\left[\max\left\{0, \hat{W}_r - \sum_{l\in J}\frac{\hat{W}_r}{\rho_r|r|}\right\}\right]^{|r|} \\
&= \frac{1}{|r|!\left(\frac{\hat{W}_r}{|r|\rho_r}\right)^{|r|}}\sum_{J\subseteq r}(-1)^{|J|}\left[\max\left\{0, \hat{W}_r - \frac{\hat{W}_r}{\rho_r|r|}\sum_{l\in J}1\right\}\right]^{|r|} \\
&= \frac{1}{|r|!\left(\frac{\hat{W}_r}{|r|\rho_r}\right)^{|r|}}\sum_{J\subseteq r}(-1)^{|J|}\left(\frac{\hat{W}_r}{|r|\rho_r}\right)^{|r|}\left[\max\left\{0, |r|\rho_r - |J|\right\}\right]^{|r|} \\
&= \frac{1}{|r|!}\sum_{J\subseteq r}(-1)^{|J|}\left[\max\left\{0, \rho_r|r| - |J|\right\}\right]^{|r|} && (12) \\
&= \frac{1}{|r|!}\sum_{i=0}^{|r|}\binom{|r|}{i}(-1)^{i}\left[\max\left\{0, \rho_r|r| - i\right\}\right]^{|r|} && (13) \\
&= \frac{1}{|r|!}\sum_{i=0}^{|r|}\frac{|r|!}{i!(|r|-i)!}(-1)^{i}\left[\max\left\{0, \rho_r|r| - i\right\}\right]^{|r|} \\
&= \sum_{i=0}^{\lfloor\rho_r|r|\rfloor}\frac{(-1)^i}{i!(|r|-i)!}(\rho_r|r| - i)^{|r|} =: g_r(\rho_r),
\end{aligned}$$

where we partitioned the sum over all subsets of legs in a route with respect to the subsets' sizes to move from (12) to (13).

Appendix B: Additional Instance Characteristics

We describe additional characteristics of the instances comprising our computational study in Section 5. Specifically, Table 7 summarizes the average shipment flow volumes (measured in weight and fractional truckloads) between the different types of origin-destination facility pairs. Next, the distribution of the allowable waiting delays for all demand commodities k, averaged across their potential routes in $\mathcal{R}_k$, is depicted in Figure 10. Finally, Figure 11 illustrates the freight transportation costs introduced in Table 2 on a 500-mile leg.

Table 7 Average volume per O-D pair across all instances.

Origins	Destinations		
	Small	Medium	Large
	LMD	LMD	LMD
Small	200 lbs	300 lbs	400 lbs
VND	0.02 TL	0.03 TL	0.03 TL
Medium	700 lbs	1,200 lbs	1,700 lbs
VND	0.06 TL	0.1 TL	0.15 TL
Large	1,400 lbs	2,300 lbs	3,700 lbs
VND	0.12 TL	0.19 TL	0.31 TL
FC	3,900 lbs	3,900 lbs	3,900 lbs
	0.33 TL	0.33 TL	0.33 TL

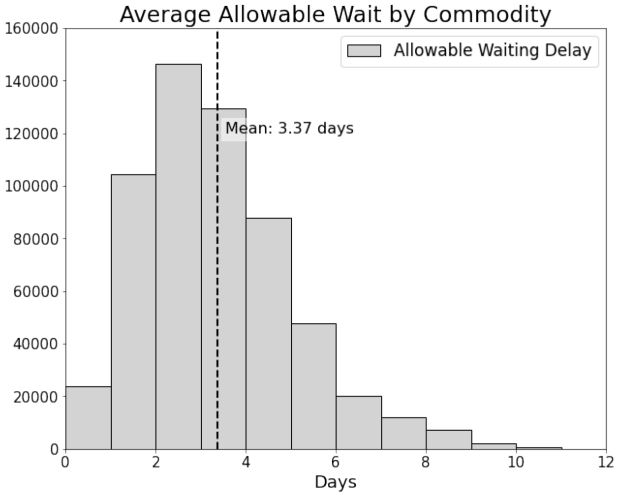


Figure 10 Distribution of average allowable waiting delays for commodities across all instances.

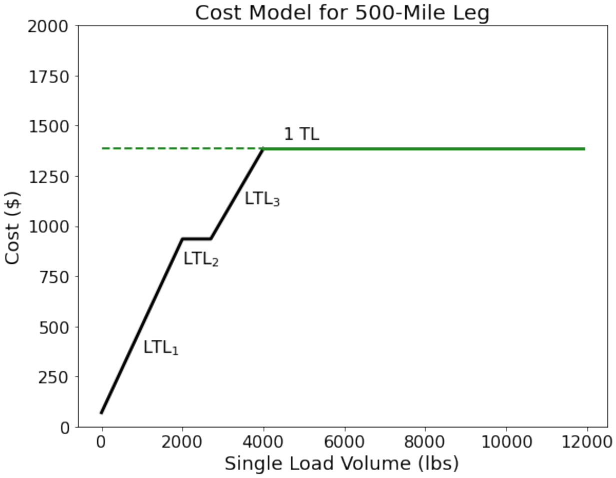


Figure 11 Freight mode costs for a 500-mile leg.

References

Andersen, J., Crainic, T.G., & Christiansen, M. 2009. Service network design with management and coordination of multiple fleets. *European Journal of Operational Research*, **193**(2), 377–389.

Bai, R., Wallace, S.W., Li, J., & Chong, A.Y.L. 2014. Stochastic service network design with rerouting. *Transportation Research Part B: Methodological*, **60**, 50–65.

Bakir, I., Erera, A.L., & Savelsbergh, M.W.P. 2021. Motor Carrier Service Network Design. *Chap. 14, pages 427–467 of:* Crainic, T.G., Gendreau, M., & Gendron, B. (eds), *Network Design with Applications to Transportation and Logistics*. Springer.

Boland, N., Hewitt, M., Marshall, L., & Savelsbergh, M.W.P. 2017. The Continuous-Time Service Network Design Problem. *Operations research*, **65**(5), 1303–1321.

Boland, N., Hewitt, M., Marshall, L., & Savelsbergh, M.W.P. 2019. The price of discretizing time: a study in service network design. *EURO Journal on Transportation and Logistics*, **8**(2), 195–216. Special Issue: Advances in vehicle routing and logistics optimization: exact methods.

Bouzaïene-Ayari, B., Gendreau, M., & Sang, N. 2001. Modeling bus stops in transit networks : A survey and new formulations: Mass public transit. *Transportation science*, **35**(3), 304–321.

Cancela, H., Mauttone, A., & Urquhart, M.E. 2015. Mathematical programming formulations for transit network design. *Transportation research. Part B: methodological*, **77**, 17–37.

Crainic, T.G. 2000. Service network design in freight transportation. *European Journal of Operational Research*, **122**(2), 272–288.

Crainic, T.G., & Rousseau, J.M. 1986. Multicommodity, multimode freight transportation: A general modeling and algorithmic framework for the service network design problem. *Transportation Research Part B: Methodological*, **20**(3), 225–242.

Crainic, T.G., & Roy, J. 1988. OR tools for tactical freight transportation planning. *European Journal of Operational Research*, **33**(3), 290–297.

Crainic, T.G., Ferland, J.A., & Rousseau, J.M. 1984. A Tactical Planning Model for Rail Freight Transportation. *Transportation Science*, **18**(2), 165–184.

Crainic, T.G., Hewitt, M., Toulouse, M., & Vu, D.M. 2016. Service Network Design with Resource Constraints. *Transportation Science*, **50**(4), 1380–1393.

Daganzo, C. 1997. *Fundamentals of transportation and traffic operations*. First edn. Emerald Group Publishing Limited.

Demir, E., Burgholzer, W., Hrušovskỳ, M., Arıkan, E., Jammernegg, W., & Van Woensel, T. 2016. A green intermodal service network design problem with travel time uncertainty. *Transportation Research Part B: Methodological*, **93**, 789–807.

Erera, A.L., Hewitt, M., Savelsbergh, M.W.P., & Zhang, Y. 2013. Improved Load Plan Design Through Integer Programming Based Local Search. *Transportation science*, **47**(3), 412–427.

Franceschi, R.D., Fischetti, M., & Toth, P. 2006. A new ILP-based refinement heuristic for vehicle routing problems. *Mathematical Programming*, **105**(2), 471–499.

Hewitt, M. 2019. Enhanced Dynamic Discretization Discovery for the Continuous Time Load Plan Design Problem. *Transportation science*, **53**(6), 1731–1750.

Hewitt, M. 2022. The Flexible Scheduled Service Network Design Problem. *Transportation Science*, **56**(4), 1000–1021.

Hewitt, M., Nemhauser, G.L., & Savelsbergh, M.W.P. 2010. Combining Exact and Heuristic Approaches for the Capacitated Fixed-Charge Network Flow Problem. *INFORMS journal on computing*, **22**(2), 314–325.

Hwang, J., Park, S., & Kong, I.Y. 2011. An integer programming-based local search for large-scale multidimensional knapsack problems. *International Journal on Computer Science and Engineering*, **3**(6), 2257–2264.

Jarrah, A., Johnson, E.L., & Neubert, L.C. 2009. Large-Scale, Less-than-Truckload Service Network Design. *Operations research*, **57**(3), 609–625.

Kang, J., Kim, S., Kim, Y., & Jang, Y.S. 2010. Generalized convolution of uniform distributions. *Journal of Applied Mathematics & Informatics*, **28**(01).

Leonard, M. 2021. *Amazon insources logistics with a growing network of delivery stations*. https://tinyurl.com/AmazonInsources, Last accessed on 2021-09-07.

Lin, C.C. 2001. The freight routing problem of time-definite freight delivery common carriers. *Transportation Research Part B: Methodological*, **35**(6), 525–547.

Lindsey, K., Erera, A.L., & Savelsbergh, M.W.P. 2016. Improved Integer Programming-Based Neighborhood Search for Less-Than-Truckload Load Plan Design. *Transportation science*, **50**(4), 1360–1379.

Lium, A.G., Crainic, T.G., & Wallace, S.W. 2009. A study of demand stochasticity in service network design. *Transportation Science*, **43**(2), 144–157.

Marshall, L., Boland, N., Savelsbergh, M.W.P., & Hewitt, M. 2021. Interval-Based Dynamic Discretization Discovery for Solving the Continuous-Time Service Network Design Problem. *Transportation science*, **55**(1), 29–51.

Mauttone, A., Cancela, H., & Urquhar, M.E. 2021. Public Transportation. *Chap. 17, pages 539–565 of:* Crainic, T.G., Gendreau, M., & Gendron, B. (eds), *Network Design with Applications to Transportation and Logistics*. Springer.

Pedersen, M.B., Crainic, T.G., & Madsen, O.B.G. 2009. Models and Tabu Search Metaheuristics for Service Network Design with Asset-Balance Requirements. *Transportation Science*, **43**(2), 158–177.

Powell, W.B. 1986. A Local Improvement Heuristic for the Design of Less-than-Truckload Motor Carrier Networks. *Transportation science*, **20**(4), 246–257.

Powell, W.B., & Koskosidis, I.A. 1992. Shipment Routing Algorithms with Tree Constraints. *Transportation Science*, **26**(3), 230–245.

Powell, W.B., & Sheffi, Y. 1983. The load planning problem of motor carriers: Problem description and a proposed solution approach. *Transportation research. Part A: general*, **17**(6), 471–480.

Savelsbergh, M.W.P., & Song, J.H. 2008. An optimization algorithm for the inventory routing problem with continuous moves. *Computers & operations research*, **35**(7), 2266–2282.

Scherr, Y.O., Saavedra, B.A. Neumann, Hewitt, M., & Mattfeld, D.C. 2019. Service network design with mixed autonomous fleets. *Transportation Research Part E: Logistics and Transportation Review*, **124**, 40–55.

Scherr, Y.O., Hewitt, M., Neumann-Saavedra, B.A., & Mattfeld, D.C. 2020. Dynamic discretization discovery for the service network design problem with mixed autonomous fleets. *Transportation Research Part B: Methodological*, **141**, 164–195.

Spiess, H., & Florian, M. 1989. Optimal strategies: A new assignment model for transit networks. *Transportation research. Part B: methodological*, **23**(2), 83–102.

The Home Depot. 2021. *Home Depot Strong - Annual Report 2020*. https://tinyurl.com/THD2020Report, Last accessed on 2021-09-07.

Wang, Z., & Qi, M. 2020. Robust service network design under demand uncertainty. *Transportation Science*, **54**(3), 676–689.

Wayfair. 2021. *Investor Presentation - Q2 2021*. https://tinyurl.com/wayfairQ22021, Last accessed on 2021-08-30.

Wieberneit, N. 2008. Service network design for freight transportation: a review. *OR Spectrum*, **30**(1), 77–112.

Zhu, E., Crainic, T.G., & Gendreau, M. 2014. Scheduled Service Network Design for Freight Rail Transportation. *Operations Research*, **62**(2), 383–400.