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 by building consolidated freight transportation loads to reduce outbound logistics cost. In this article, we study a middle mile network design problem with fixed origins and destinations to determine minimum cost consolidation plans that 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 difficult-to-solve 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. In most e-commerce networks, goods are held in inventory at one or more fulfillment centers (FCs) and then are shipped in response to orders. E-retailers may also coordinate shipments to customers directly from product vendors. Large e-retailers who operate such networks, and especially those that coordinate shipments both from FCs and also directly

* Corresponding author.
from vendors, can generate substantial cost savings by building and operating private middle mile consolidation networks. However, the design of such networks can be challenging.

In this paper, we refer to a middle mile network as the outbound freight transportation network operated to move goods ordered by customers from origin stocking locations into destination facilities where shipments are transferred to last mile distribution (LMD) partners; such partners may include package express 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. Importantly, a middle mile network must ensure that shipments can be delivered to customers by their promised delivery times, and these lead-time promises have tightened as competition between retailers has increased. Some retailers do not use a middle mile network and hand off shipments to LMD partners at or near origin facilities. However, significant cost savings may result when retailers find opportunities to create consolidated loads from origins to facilities much closer to their final delivery destinations. These loads are typically outsourced to truckload or LTL carriers for transportation.

Amazon, Wayfair, and The Home Depot are current examples of U.S. shippers who actively manage complex middle mile networks for e-commerce fulfillment. Amazon was one of the first to focus on so-called middle mile in-sourcing. They continue to grow their set of subsidiary LMD centers (i.e., delivery stations) which now includes 1,500 worldwide locations (Leonard, M. 2021); in addition to these centers, other middle mile destinations include facilities operated by non-subsidiary LMD carrier partners. Amazon now operates branded truckload trailer equipment in their middle mile network, using trucking carrier partners to ensure capacity availability and to allow efficient drop-and-hook operations (Martineau, P. 2020). 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 their private middle mile network (Wayfair 2021). As a final example, The Home Depot is expanding a middle mile fulfillment system as part of their One Supply Chain initiative (The Home Depot 2021), leveraging new and existing fulfillment center locations as consolidation locations. Each of these companies seeks to speed up order delivery to customers and decrease reliance on third party delivery companies, both for cost and reliability.

In this paper, we develop planning models for tactical middle mile network design problems where shipment origin and destination facilities are given. That is, each forecasted customer order is assumed to be fulfilled via a shipment from a specific origin facility (vendor or FC) to a specific destination facility. Suppose that each customer shipment has a lead-time requirement that specifies the maximum amount of time allowed to transfer from origin to destination. Middle mile planning models specify how to move shipments from origins to destinations in consolidated loads. Loads
may be direct, consolidating shipments from a single origin facility into a destination LMD facility. A network with only direct loads does not require any shipment sorting except at the origins. To reduce costs and/or lead times, we consider networks that also allow shipments to be transferred between loads at intermediate sorting facilities; such transfers may occur at cross-docks or sortation hubs, at FCs, or at other distribution facilities operated by the retailer. Thus, planning models specify a path for each shipment from its origin to destination, either direct or through intermediate sorting facilities. To ensure the feasibility of the plan and estimate its required transportation cost, models also determine the frequency of loads required to be moved between pairs of facilities. Load frequencies determine whether enough capacity is available to accommodate all shipments. Shipments also must arrive at their destinations within their lead-time requirements, and the total middle mile network operating cost including transportation and sorting costs should be minimized.

Traditional flat network service network design (SND) optimization models for trucking were developed to configure consolidation networks operated by LTL carriers (see the review in Bakir et al., 2021), and adaptations of these models can be used for middle mile network planning. 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 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. E-commerce customers are now promised tighter delivery lead times, and middle mile consolidation plans must be designed to meet these customer expectations. One approach to do so would be to use a detailed time-expanded network model that schedules load dispatches and freight transfers at specific times during a planning horizon to ensure planned shipments arrive on time. Such models may not be necessary or appropriate for creating tactical plans using forecasted average demand volumes. Additionally, real-world instances of such models may be massive for larger retailers and difficult to optimize exactly or suboptimally via heuristics. We propose a different methodology for tactical planning that can serve as a useful alternative to such detailed models.

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 planning models more 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 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;
– show the effectiveness of the MMCW model for solving small- to medium-sized middle mile network design problems with lead-time constraints when embedded within a practical integer-programming-based (IP-based) local search heuristic;
– 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;
– 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. To help ensure that shipments between terminal pairs met desired customer service levels, initial models specified minimum weekly truckload frequencies on arcs with positive truck flows to control waiting delays for
transferred shipments (Powell & Sheffi 1983, Powell 1986, Powell & Koskosidis 1992). A different stream of research proposed flat network models for rail freight SND tactical planning applications (Crainic et al. 1984, Crainic & Rousseau 1986) and then modified 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; since the resulting integer programming models were intractable, a heuristic solution approach is proposed for iteratively updating integer service frequencies. We propose a similar approach to model waiting delays as a function of service frequency, but instead of penalizing delay we create probabilistic constraints to ensure that shipments reach destinations within the promised lead time and attempt to solve the resulting integer programming models exactly or to within a provable optimality gap.

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). 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). Other variants of SSND problems have also 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). 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 at any location 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 will adopt a similar idea in this paper.

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. IP-based local search is a MIP solution approach that combines exact and heuristic approaches (Hewitt et al. 2010). In this framework, a restricted version of the full MIP is solved at each iteration (Erera et al. 2013, Lindsey et al. 2016) in an attempt to improve an incumbent solution. We utilize this general framework in our work by optimizing consolidation in small parts of the network iteratively.

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 provide sufficient freight transportation capacity between its facilities to satisfy shipment demand and lead-time constraints. 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, where a load is a consolidated set of shipments to be dispatched along a leg at a single point in time; these intermediate locations may be FCs, cross-docks, or other transfer terminals. Each location \( i \in \mathcal{N} \) 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. 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 \( Q^{\text{max}}_{lm} \) and lower bound \( Q^{\text{min}}_{lm} \) on the size of each dispatched load.

For middle mile networks, typical transportation types include truckload and LTL trucking. The load lower and upper bounds are used to model both physical constraints on load size by mode but also 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 \( R := \bigcup_{k \in \mathcal{K}} R_k \) 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 \((l, m)\) per time (e.g., 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 \( r \in R \) 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 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 \( r \in R_k \) 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 \( R \) 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:

\[
\begin{align}
\min_{x, y, f, v} & \quad \sum_{r \in R} C_r x_r + \sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}_l} (A_{lm} f_{lm} + B_{lm} v_{lm}) \\
\text{s.t.} & \quad \sum_{r \in R_k} x_r = 1, \quad \forall k \in \mathcal{K}, \quad (1a) \\
& \quad \sum_{m \in \mathcal{M}_l} v_{lm} = \sum_{k \in \mathcal{K}} \sum_{(r \in R_k \mid r \ni l)} V_k x_r, \quad \forall l \in \mathcal{L}, \quad (1b) \\
& \quad v_{lm} \leq Q_{lm}^{\text{max}} f_{lm}, \quad \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l, \quad (1c) \\
& \quad v_{lm} \geq Q_{lm}^{\text{min}} f_{lm}, \quad \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l, \quad (1d)
\end{align}
\]
\[ \sum_{m \in M_l} y_{lm} \leq 1, \quad \forall l \in \mathcal{L}, \quad (1f) \]
\[ f_{lm} \leq F_{lm} y_{lm}, \quad \forall l \in \mathcal{L}, \forall m \in M_l, \quad (1g) \]
\[ x_r \in \{0, 1\}, \quad \forall r \in \mathcal{R}, \quad (1h) \]
\[ y_{lm} \in \{0, 1\}, \quad \forall l \in \mathcal{L}, \forall m \in M_l, \quad (1i) \]
\[ f_{lm} \in \mathbb{Z}_{\geq 0}, \quad \forall l \in \mathcal{L}, \forall m \in M_l, \quad (1j) \]
\[ v_{lm} \geq 0, \quad \forall l \in \mathcal{L}, \forall m \in M_l. \quad (1k) \]

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. 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 that a minor assumption of this model, consistent with nearly all flow and load planning models in the literature, is that shipments can be fluidly packed into discrete loads without considering discrete bin packing considerations. However, a greater 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. If the shipments for each commodity \( k \) are then assumed to arrive at \( o_k \) according
to a uniform distribution during any headway, it is reasonable to assume further that all arriving shipments to be dispatched from a facility $h$ (whether originating or transferring) will have similarly distributed arrival times. Thus, the waiting delay experienced by any individual shipment on each leg $l$ can be modeled as a uniform random variable $W_l \sim \text{Uniform}(0, \frac{1}{f_l})$.

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. Thus, 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 $r$, denoted $\hat{W}_r$, if and only if the total waiting delay along that route does not exceed $\hat{W}_r$. This can be expressed as $\sum_{l \in r} W_l \leq \hat{W}_r$, which involves the random variables $W_l$. One method that aims to capture the variability and ensure this condition is satisfied with a certain probability $p$ is to consider the following chance constraint:

$$\mathbb{P}\left(\sum_{l \in r} W_l \leq \hat{W}_r \right) \geq p.$$  \hfill (2)

Constraint (2) 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. Since the waiting delay experienced by shipments to be dispatched on leg $l$ is assumed to be given by $W_l \sim \text{Uniform}(0, \frac{1}{f_l})$, the probability that the commodity traveling along $r$ arrives on time to its destination is given by the following expression (Kang et al. 2010):

$$\mathbb{P}\left(\sum_{l \in r} W_l \leq \hat{W}_r \right) = \frac{1}{|r|!} \frac{1}{\prod_{l \in r} f_l} \sum_{J \subseteq r} (-1)^{|J|} \left[ \max \left\{ 0, \hat{W}_r - \sum_{l \in J} \frac{1}{f_l} \right\} \right]^{|r|}.$$  \hfill (3)

However, the resulting constraint (2) is nonlinear in the load dispatch frequencies and cannot be included directly in the optimization model (1). Instead, we approximate (2) using integer linear constraints and add them to model (1).

The first step consists of approximating constraint (2) 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 (2) is equivalent to:

$$\sum_{l \in r} \frac{1}{2} \frac{1}{f_l} \leq \hat{W}_r.$$  \hfill (4)

Similarly, when the probability of on-time arrival is $p = 1$, chance constraint (2) is equivalent to:

$$\sum_{l \in r} \frac{1}{f_l} \leq \hat{W}_r.$$  \hfill (5)
From these observations, given a general on-time arrival probability $p$, we approximate chance constraint (2) using the following constraint:

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

where $c_r \in [0,1]$ is a conservatism hyperparameter that can be adjusted to guarantee the probability of meeting the commodity lead-time requirement $p$ desired by the shipper.

Given deterministic and homogeneous headways, setting $c_r = 0.5$ (respectively $c_r = 1$) 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 $c_r$ given the desired on-time arrival probability $p$ is challenging. A low value of $c_r$ will allow the selection of load plans that will not meet the on-time arrival probability, while a high value of $c_r$ 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 $c_r$ for which constraints (6) guarantee that a commodity traveling on route $r$ meets the lead-time requirement with probability $p$. This problem can be formulated as follows:

$$\min c_r = \frac{\hat{W}_r}{\sum_{l \in r} \frac{1}{f_l}}$$

s.t. $\mathbb{P}\left(\sum_{l \in r} W_l \leq \hat{W}_r\right) \geq p, \quad \forall f \in \mathbb{Z}_{>0}^r \mid c_r \sum_{l \in r} \frac{1}{f_l} \leq \hat{W}_r,$$

$\forall f' \in \mathbb{Z}_{>0}^r.$

Equivalently, we select for each route $r$ the lowest value of $c_r$ 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 $c_r$ value that separates all combinations of load dispatch frequencies that satisfy chance constraint (2) from the ones that do not. Thus, this approach can lead to setting $c_r$ 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 (6) 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 (6) 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 \(f \in \mathcal{F}_{lm} := \{1, \ldots, F_{lm}\}\) satisfying the maximum load dispatch frequency \(F_{lm}\), we define the binary variable \(z_{lmf}\). We then substitute the frequency variables as follows:

\[
f_{lm} = \sum_{f \in \mathcal{F}_{lm}} f z_{lmf}, \quad \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l.
\]

We can now reformulate (6) as linear integer constraints, thus leading to the following formulation of the MMCW model:

\[
\begin{align*}
\min_{x, z, v} \quad & \sum_{r \in \mathcal{R}} C_r x_r + \sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}_l} \left[ A_{lm} \left( \sum_{f \in \mathcal{F}_{lm}} f z_{lmf} \right) + B_{lm} v_{lm} \right] \\
\text{s.t.} \quad & \sum_{r \in \mathcal{R}} x_r = 1, \quad \forall k \in \mathcal{K}, \quad (8a) \\
& \sum_{m \in \mathcal{M}_l} v_{lm} = \sum_{k \in \mathcal{K}} \left( \sum_{r \in \mathcal{R}_k, r \ni l} V_k x_r \right), \quad \forall l \in \mathcal{L}, \quad (8b) \\
& v_{lm} \leq Q_{lm}^{\text{max}} \sum_{f \in \mathcal{F}_{lm}} f z_{lmf}, \quad \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l, \quad (8c) \\
& v_{lm} \geq Q_{lm}^{\text{min}} \sum_{f \in \mathcal{F}_{lm}} f z_{lmf}, \quad \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l, \quad (8d) \\
& \sum_{m \in \mathcal{M}_l} \sum_{f \in \mathcal{F}_{lm}} z_{lmf} \leq 1, \quad \forall l \in \mathcal{L}, \quad (8e) \\
& c_r \sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}_l} \sum_{f \in \mathcal{F}_{lm}} \frac{1}{f} z_{lmf} \leq \hat{W}_r x_r + c_r |r| (1 - x_r), \quad \forall r \in \mathcal{R}, \quad (8f) \\
& x_r \in \{0, 1\}, \quad \forall r \in \mathcal{R}, \quad (8g) \\
& z_{lmf} \in \{0, 1\}, \quad \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l, \forall f \in \mathcal{F}_{lm}, \quad (8h) \\
& v_{lm} \geq 0, \quad \forall l \in \mathcal{L}, \forall m \in \mathcal{M}_l. \quad (8i)
\end{align*}
\]

Objective (8a) and constraints (8d)-(8e) are obtained by applying the variable replacement (7) to objective (1a) and constraints (1d)-(1e), respectively. Constraints (8f) replace constraints (1f)-(1g) and select at most one frequency for each lane. Constraints (8g) provide a linear formulation of constraints (6) 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 \( z_{lmf} \), \( \forall (l,m,f) \in \mathcal{L} \times \mathcal{M}_l \times \mathcal{F}_{lm} \) 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 leads to an upper bound on the optimal solution value to the MMCW problem.

To avoid introducing additional binary variables, we now develop a simpler alternative for finding a reasonable solution to the MMCW problem (and an associated upper bound on its optimal value). This approach restricts the space of feasible solutions by allocating fixed fractions of a route’s total allowable waiting delay \( \text{a priori} \) 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 \emph{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 (2) using the following hyperparameterized constraint on every leg \( l \) of a selected route \( r \):

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

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

\[
\sum_{m \in \mathcal{M}_l} f_{lm} \geq c_r \frac{|r|}{\hat{W}_r} x_r, \quad \forall r \in \mathcal{R}, \forall l \in r.
\]  

(9)

The right-hand side of constraints (9) 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 \( c_r \). 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 \textit{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 (9) and the allowable waiting delay allocated to \( l_2 \).

Similarly to the MMCW model, we aim to determine for every selected route \( r \in \mathcal{R} \) the lowest hyperparameter \( c_r \) that will guarantee that the load plans satisfying constraints (9) meet the corresponding lead-time requirement with probability at least \( p \). This problem can be formulated as follows:

\[
\begin{align*}
\min c_r &= \frac{f' \hat{W}_r}{|r|} \\
\text{s.t. } P \left( \sum_{l \in r} W_l \leq \hat{W}_r \right) &\geq p, \quad \forall f \in \mathbb{Z}_{>0}^r \quad \left| \frac{1}{f_l} - \frac{\hat{W}_r}{c_r |r|} \right| \forall l \in r, \\
f' &\in \mathbb{Z}_{>0}.
\end{align*}
\]

Interestingly, we can derive the following result for the load dispatch frequencies satisfying constraints (9). Specifically, for every route \( r \in \mathcal{R} \) and every set of frequencies \( f \in \mathbb{Z}_{>0}^r \) satisfying \( \frac{1}{f_l} \leq \frac{\hat{W}_r}{c_r |r|} \) for every \( l \in r \), we have:

\[
P \left( \sum_{l \in r} W_l \leq \hat{W}_r \right) = \frac{1}{|r|!} \sum_{J \subseteq r} (-1)^{|J|} \left[ \max \left\{ 0, \hat{W}_r - \sum_{l \in J} \frac{\hat{W}_r}{c_r |r|} \right\} \right]^{\left| r \right|} \\
= \frac{1}{|r|!} \sum_{i=0}^{|r|} \binom{|r|}{i} (-1)^i \left[ \max \left\{ 0, c_r |r| - i \right\} \right]^{\left| r \right|} \\
= \sum_{i=0}^{|r|} \binom{|r|}{i} (-1)^i \left( c_r |r| - i \right)^{|r|} = g_r(c_r).
\]

Thus, given a desired on-time probability \( p \), we can determine the corresponding conservatism level \( c_r \) for each constraint (9) 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 \( c_r \) by solving the polynomial equation \( g_r(x) - p = 0 \) on each interval \( \left[ \frac{j-1}{|r|}, \frac{j}{|r|} \right], \quad j \in \{1, \ldots, |r|\} \).

Surprisingly, we observe that computing the conservatism level \( c_r \) 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 1 illustrates the guaranteed on-time probability as a function of the conservatism level and the number of legs for this model.

Again, we find that when the probability of on-time arrival is $p = 0.5$ (resp. $p = 1$), setting the hyperparameter to $c_r = 0.5$ (resp. $c_r = 1$) ensures that the load plans that satisfy constraints (9) 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 $c_r > 0.5$, the probability of on-time arrival increases with the number of legs $|r|$. On the other hand, if $c_r < 0.5$, 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 $(\sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}_l} (|F_{lm}| - 1))$ 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 $\alpha |R|$ routes are included in $R^{(t)}$, 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 $R^{(t)}$. However, Algorithm 2 selects $o_s$ from a subset of origin facilities $O_h$ 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 $o_s \in O_h$ 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 $O_h$. 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 $R^{(t)}$ 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 $R^{(t)}$

1. Set $R^{(t)} \leftarrow \emptyset$;
2. Set $\mathcal{O} \leftarrow \{o_k, \ \forall k \in \mathcal{K}\}$;
3. Set $\hat{V} \leftarrow \sum_{k \in \mathcal{K}} V_k$;
4. **while** $|R^{(t)}| < \alpha |\mathcal{R}| \text{ and } \mathcal{O} \neq \emptyset$ **do**
   5. Set $w(o) \leftarrow \frac{1}{\hat{V}} \sum_{k \in \mathcal{K} | o_k = o} V_k,$ $\forall o \in \mathcal{O}$;
   6. Select origin $o_s$ randomly from $\mathcal{O}$ using probability mass function $w$;
   7. $R^{(t)} \leftarrow R^{(t)} \cup (\bigcup_{k \in \mathcal{K} | o_k = o_s} \mathcal{R}_k)$;
   8. $\mathcal{O} \leftarrow \mathcal{O} \setminus \{o_s\}$;
   9. $\hat{V} \leftarrow \hat{V} - \sum_{k \in \mathcal{K} | o_k = o_s} V_k$;
5. **end**
11. return $R^{(t)}$

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,
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 $\mathcal{R}^{(t)}$

1. Set $\mathcal{R}^{(t)} \leftarrow \emptyset$;
2. Set $\mathcal{K}_h \leftarrow \{k \in \mathcal{K} \mid \text{at least one route } r \in \mathcal{R}_k \text{ includes location } h \text{ as a transfer point}\}$;
3. Set $\mathcal{O}_h \leftarrow \{o_k, \forall k \in \mathcal{K}_h\}$;
4. Set $\hat{V} \leftarrow \sum_{k \in \mathcal{K}_h} V_k$;
5. **while** $|\mathcal{R}^{(t)}| < \alpha |\mathcal{R}|$ and $\mathcal{O}_h \neq \emptyset$ do
6. \[\text{Set } w(o) \leftarrow \frac{1}{\hat{V}} \sum_{k \in \mathcal{K}_h | o_k = o} V_k, \forall o \in \mathcal{O}_h;\]
7. \[\text{Select origin } o_s \text{ randomly from } \mathcal{O} \text{ using probability mass function } w;\]
8. \[\mathcal{R}^{(t)} \leftarrow \mathcal{R}^{(t)} \cup (\cup_{k \in \mathcal{K}_h | o_k = o_s} \mathcal{R}_k);\]
9. \[\mathcal{O}_h \leftarrow \mathcal{O}_h \setminus \{o_s\};\]
10. \[\hat{V} \leftarrow \hat{V} - \sum_{k \in \mathcal{K}_h | o_k = o_s} V_k;\]
11. end
12. **return** $\mathcal{R}^{(t)}$

the incumbent is updated. Within each iteration of the search, the incumbent solution is used as a warm-start solution. 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.

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 $T_{\text{MMCW}}$ spent improving the solution using the complete MMCW model.
Algorithm 3: Single-Phase IP-Based Local Search

**Input:** MIP, initial feasible solution \((\hat{x}, \hat{v}, (\hat{f}, \hat{y}) \text{ or } \hat{z})\), objective value \(\hat{w}\), and \(\text{list}_H\) as an ordered list of transfer locations \(N_H\)

**Result:** Improved feasible solution and improved objective value

1. Set \(\text{val} \leftarrow \hat{w}\), \(\text{T}_{\text{run}} \leftarrow 0\), \(\text{iter} \leftarrow 0\), \(\text{neighborhood\_select} \leftarrow 1\), and \(i \leftarrow 1\);
2. while \(\text{T}_{\text{run}} \leq T\) do
   3. if \(\text{neighborhood\_select} = 1\) then
      4. Select \(\mathcal{R}^{(t)}\) using Algorithm 1;
   else
      5. \(h \leftarrow \text{list}_H[i];\)
      6. Select \(\mathcal{R}^{(t)}\) using Algorithm 2 and \(h\) as selected intermediate facility;
      7. if \(i < |\text{list}_H|\) then
         8. \(i \leftarrow i + 1;\)
      else
         9. \(i \leftarrow 1;\)
   10. Solve MIP after adding constraints \(x_r = \hat{x}_r\), \(\forall r \in \mathcal{R}\setminus\mathcal{R}^{(t)}\), using \((\hat{x}, \hat{v}, (\hat{f}, \hat{y}) \text{ or } \hat{z})\) as warm-start solution;
   11. \(T_{\text{MIP}} \leftarrow \text{MIP solving time};\)
   12. \(\text{newval} \leftarrow \text{MIP solution's objective value};\)
   13. if \(\text{newval} < \text{val}\) then
      14. Set \((\hat{x}, \hat{v}, (\hat{f}, \hat{y}) \text{ or } \hat{z}) \leftarrow \text{MIP solution};\)
      15. Set \(\text{val} \leftarrow \text{newval}, \text{iter} \leftarrow 0;\)
   else
      16. Set \(\text{iter} \leftarrow \text{iter} + 1;\)
   17. if \(\text{iter} = N\) then
      18. if \(\text{neighborhood\_select} = 1\) then
         19. \(\text{neighborhood\_select} \leftarrow 2;\)
      else
         20. \(\text{neighborhood\_select} \leftarrow 1;\)
      21. Set \(\text{iter} \leftarrow 0;\)
      22. \(\text{T}_{\text{run}} \leftarrow \text{T}_{\text{run}} + T_{\text{MIP}};\)
   end
23. return \((\hat{x}, \hat{v}, (\hat{f}, \hat{y}) \text{ or } \hat{z}), \text{val}\)

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

1. Set \(\hat{f}_{lm} \leftarrow \sum_{f \in F_{lm}} f_{lmf} \quad \forall l \in L, \forall m \in M_l\);
2. Set \(\hat{y}_{lm} \leftarrow \sum_{f \in F_{lm}} \hat{z}_{lmf} \quad \forall l \in L, \forall m \in M_l\);
3. Set \(\hat{z} \leftarrow \hat{w}\);
4. Run single-phase IPBLS (Algorithm 3) using MMCW-A model for \(T_{MMCW-A}\) time with initial solution \((\hat{x}, \hat{v}, (\hat{f}, \hat{y}))\) and objective value \(\hat{w}\) as input;
5. Set \((x', v', (f', y')) \leftarrow\) output solution;
6. Set \(\hat{w} \leftarrow\) output objective value;
7. Set \(z_{lmf} \leftarrow 1\{f_{lmf} = f\} \quad \forall l \in L, \forall m \in M_l, \forall f \in F_{lm}\);
8. Run single-phase IPBLS (Algorithm 3) using MMCW model for \(T_{MMCW}\) time with initial solution \((x', v', z')\) and objective value \(\hat{w}\) as input;
9. Set \((\hat{x}, \hat{v}, \hat{z}) \leftarrow\) output solution;
10. Set \(\hat{w} \leftarrow\) output objective value;
11. return \((\hat{x}, \hat{v}, \hat{z}), \hat{w}\)

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 that are not described in more detail in this paper. When selecting the subsets of routes \(R^{(d)}\) 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. It may be useful in practice to have \(\alpha\) vary as Algorithm 3 proceeds; larger values of \(\alpha\) can be used in later iterations to intensify the search. The number of non-improving neighborhood searches allowed before switching neighborhood selection methods is set to \(N = 5\). We experimented with slightly different values 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, and a smaller value of $N$ provides the heuristic with enough time to use both randomized neighborhoods multiple times.

### 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. Figure 2 shows the vendor, FC, and LMD locations for Group 4 - Instance 1; FC locations are identical in all groups and instances. See Appendix A for additional details on the instances, including characteristics of the average flow between location types.

![Figure 2](locations_group4_instance1.png)

**Figure 2** Locations for Group 4 - Instance 1.

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. Estimates of freight mode costs were derived using actual costs provided again by our
Table 1  Instance characteristics.

<table>
<thead>
<tr>
<th>Gr</th>
<th>Locations / Average Volume in Truckloads</th>
<th>Comm.</th>
<th>Inst.</th>
<th>Lanes</th>
<th>Routes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small VND</td>
<td>Medium VND</td>
<td>Large VND</td>
<td>FC</td>
<td>Small LMD</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>30</td>
<td>15</td>
<td>8</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>40</td>
<td>20</td>
<td>8</td>
<td>50</td>
</tr>
<tr>
<td>7</td>
<td>150</td>
<td>50</td>
<td>25</td>
<td>8</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>250</td>
<td>75</td>
<td>40</td>
<td>8</td>
<td>70</td>
</tr>
<tr>
<td>9</td>
<td>375</td>
<td>100</td>
<td>50</td>
<td>8</td>
<td>80</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Mode</th>
<th>Weight (lbs)</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>TL</td>
<td>$0 &lt; w \leq 12,000$</td>
<td>$750 + 1.27d$</td>
</tr>
<tr>
<td>LTL_1</td>
<td>$0 &lt; w \leq 2,000$</td>
<td>$0.05(750 + 1.27d) + w(0.234 + 0.0004d)$</td>
</tr>
<tr>
<td>LTL_2</td>
<td>$2,000 &lt; w \leq 2,700$</td>
<td>$0.05(750 + 1.27d) + 2000(0.234 + 0.0004d)$</td>
</tr>
<tr>
<td>LTL_3</td>
<td>$2,700 &lt; w \leq 4,000$</td>
<td>$0.8w(0.234 + 0.0004d)$</td>
</tr>
</tbody>
</table>

Table 2 Freight mode costs for a single load.

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 $R_k$ with the 5 most operationally reasonable route options for each commodity $k$. The direct, single-leg route connecting the commodity origin to destination is always included. 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. 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 $c_r = 0.5$ (to meet lead-time requirements in expectation) in the computational experiments to follow, except those in Section 5.3 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 \( L \), substantially reducing the number of decision variables and related constraints.

5.2. The Effect of Constraining Lead Time

First, we study the effect of adding lead-time constraints when optimizing a middle mile network design using the MMCW model. We present in Table 3 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.

<table>
<thead>
<tr>
<th>Group</th>
<th>MMC</th>
<th>MMCW</th>
<th>MMC</th>
<th>MMCW</th>
<th>MMC</th>
<th>MMCW</th>
<th>MMC</th>
<th>MMCW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IPBLS 12-hr</td>
<td>Vol-Wtd Route</td>
<td>Vol-Wtd Route</td>
<td>Avg Load Disp Freq (#/week)</td>
<td>Loads/Week</td>
<td>Vol-Wtd Utilization</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Objective</td>
<td>Length</td>
<td>Length</td>
<td>LTL</td>
<td>TL</td>
<td>LTL</td>
<td>TL</td>
<td>LTL</td>
</tr>
<tr>
<td>Group 1</td>
<td>$90,000</td>
<td>1.5</td>
<td>0.3</td>
<td>1.0</td>
<td>1.1</td>
<td>30</td>
<td>45</td>
<td>80%</td>
</tr>
<tr>
<td>MMC</td>
<td>$118,000</td>
<td>1.5</td>
<td>0.3</td>
<td>1.3</td>
<td>1.9</td>
<td>70</td>
<td>40</td>
<td>67%</td>
</tr>
<tr>
<td>MMCW</td>
<td>$118,000</td>
<td>1.5</td>
<td>0.3</td>
<td>1.3</td>
<td>1.9</td>
<td>70</td>
<td>40</td>
<td>67%</td>
</tr>
<tr>
<td>Group 2</td>
<td>$256,000</td>
<td>1.6</td>
<td>0.4</td>
<td>1.0</td>
<td>1.2</td>
<td>20</td>
<td>140</td>
<td>83%</td>
</tr>
<tr>
<td>MMC</td>
<td>$311,000</td>
<td>1.7</td>
<td>0.4</td>
<td>1.1</td>
<td>1.8</td>
<td>70</td>
<td>170</td>
<td>77%</td>
</tr>
<tr>
<td>MMCW</td>
<td>$311,000</td>
<td>1.7</td>
<td>0.4</td>
<td>1.1</td>
<td>1.8</td>
<td>70</td>
<td>170</td>
<td>77%</td>
</tr>
<tr>
<td>Group 3</td>
<td>$573,000</td>
<td>1.7</td>
<td>0.4</td>
<td>1.0</td>
<td>1.3</td>
<td>30</td>
<td>310</td>
<td>88%</td>
</tr>
<tr>
<td>MMC</td>
<td>$693,000</td>
<td>1.8</td>
<td>0.4</td>
<td>1.2</td>
<td>2.0</td>
<td>140</td>
<td>370</td>
<td>82%</td>
</tr>
<tr>
<td>MMCW</td>
<td>$693,000</td>
<td>1.8</td>
<td>0.4</td>
<td>1.2</td>
<td>2.0</td>
<td>140</td>
<td>370</td>
<td>82%</td>
</tr>
<tr>
<td>Group 4</td>
<td>$1,249,000</td>
<td>1.7</td>
<td>0.3</td>
<td>1.0</td>
<td>1.4</td>
<td>60</td>
<td>680</td>
<td>91%</td>
</tr>
<tr>
<td>MMC</td>
<td>$1,522,000</td>
<td>2.0</td>
<td>0.4</td>
<td>1.4</td>
<td>2.4</td>
<td>230</td>
<td>843</td>
<td>85%</td>
</tr>
<tr>
<td>MMCW</td>
<td>$1,522,000</td>
<td>2.0</td>
<td>0.4</td>
<td>1.4</td>
<td>2.4</td>
<td>230</td>
<td>843</td>
<td>85%</td>
</tr>
<tr>
<td>Group 5</td>
<td>$1,888,000</td>
<td>1.7</td>
<td>0.3</td>
<td>1.0</td>
<td>1.5</td>
<td>120</td>
<td>1,060</td>
<td>91%</td>
</tr>
<tr>
<td>MMC</td>
<td>$2,273,000</td>
<td>2.0</td>
<td>0.4</td>
<td>1.6</td>
<td>2.7</td>
<td>400</td>
<td>1,300</td>
<td>86%</td>
</tr>
<tr>
<td>MMCW</td>
<td>$2,273,000</td>
<td>2.0</td>
<td>0.4</td>
<td>1.6</td>
<td>2.7</td>
<td>400</td>
<td>1,300</td>
<td>86%</td>
</tr>
<tr>
<td>Group 6</td>
<td>$2,956,000</td>
<td>1.8</td>
<td>0.2</td>
<td>1.0</td>
<td>1.7</td>
<td>60</td>
<td>1,300</td>
<td>92%</td>
</tr>
<tr>
<td>MMC</td>
<td>$3,730,000</td>
<td>1.9</td>
<td>0.4</td>
<td>1.9</td>
<td>3.2</td>
<td>790</td>
<td>2,110</td>
<td>85%</td>
</tr>
<tr>
<td>MMCW</td>
<td>$3,730,000</td>
<td>1.9</td>
<td>0.4</td>
<td>1.9</td>
<td>3.2</td>
<td>790</td>
<td>2,110</td>
<td>85%</td>
</tr>
<tr>
<td>Group 7</td>
<td>$3,862,000</td>
<td>1.8</td>
<td>0.2</td>
<td>1.0</td>
<td>1.7</td>
<td>60</td>
<td>2,230</td>
<td>93%</td>
</tr>
<tr>
<td>MMC</td>
<td>$5,066,000</td>
<td>2.0</td>
<td>0.4</td>
<td>2.0</td>
<td>3.4</td>
<td>1,320</td>
<td>2,830</td>
<td>81%</td>
</tr>
<tr>
<td>MMCW</td>
<td>$5,066,000</td>
<td>2.0</td>
<td>0.4</td>
<td>2.0</td>
<td>3.4</td>
<td>1,320</td>
<td>2,830</td>
<td>81%</td>
</tr>
<tr>
<td>Group 8</td>
<td>$6,575,000</td>
<td>1.9</td>
<td>0.1</td>
<td>1.0</td>
<td>2.1</td>
<td>50</td>
<td>3,860</td>
<td>94%</td>
</tr>
<tr>
<td>MMC</td>
<td>$10,938,000</td>
<td>2.1</td>
<td>0.3</td>
<td>2.5</td>
<td>6.3</td>
<td>2,000</td>
<td>6,020</td>
<td>61%</td>
</tr>
<tr>
<td>MMCW-A+MMCW</td>
<td>$10,938,000</td>
<td>2.1</td>
<td>0.3</td>
<td>2.5</td>
<td>6.3</td>
<td>2,000</td>
<td>6,020</td>
<td>61%</td>
</tr>
<tr>
<td>Group 9</td>
<td>$9,230,000</td>
<td>1.9</td>
<td>0.1</td>
<td>1.0</td>
<td>2.3</td>
<td>100</td>
<td>5,490</td>
<td>93%</td>
</tr>
<tr>
<td>MMC</td>
<td>$19,730,000</td>
<td>2.2</td>
<td>0.3</td>
<td>2.5</td>
<td>7.7</td>
<td>2,610</td>
<td>10,310</td>
<td>41%</td>
</tr>
<tr>
<td>MMCW-A+MMCW</td>
<td>$19,730,000</td>
<td>2.2</td>
<td>0.3</td>
<td>2.5</td>
<td>7.7</td>
<td>2,610</td>
<td>10,310</td>
<td>41%</td>
</tr>
</tbody>
</table>

Table 3 Comparison of the MMC and MMCW model solutions.
As expected, we observe 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 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 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 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.

![Comparison of FC-to-FC truckload lane volume and load dispatch frequencies for the MMC and MMCW models (Group 4 - Instance 1).](image)

It is interesting to note that when lead-time constraints are enforced, we observe an increase in the 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 3, 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 4 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.

Finally, we show two examples from the Group 4 instances in Figure 5 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 5](image)

**Figure 5** 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.

### 5.3. 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 $c_r$ 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. To measure service level in a solution, we calculate the volume-weighted expected on-time probability (vOTP) of a solution as follows:

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

where the on-time probability of an individual commodity, $P\left(\sum_{l \in r} W_l \leq \hat{W}_r\right)$ when using route $r$, is calculated using (3), the assigned load dispatch frequencies $f_l \forall l \in r$ from the solution, and the allowable waiting delay $\hat{W}_r$. 
The results for the two mid-sized groups with four different values of on-time probability guarantees are shown in Table 4, as well as results for the associated MMC solution. 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.

<table>
<thead>
<tr>
<th>Group</th>
<th>p</th>
<th>Model</th>
<th>12-hr IPBLS Obj</th>
<th>vOTP</th>
<th>Vol-Wtd Route Length</th>
<th>Avg Load Disp Freq (#/week)</th>
<th>Loads/Week Vol-Wtd Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>MMC</td>
<td>$1,888,000</td>
<td>0.47</td>
<td>1.7</td>
<td>1.0 1.5</td>
<td>120 1,060 91%</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>MMCW</td>
<td>$2,273,000</td>
<td>0.79</td>
<td>2.0</td>
<td>1.6 2.7</td>
<td>400 1,300 86%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$2,587,000</td>
<td>0.93</td>
<td>2.1</td>
<td>1.9 3.9</td>
<td>700 1,430 82%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>MMCW</td>
<td>$2,463,000</td>
<td>0.92</td>
<td>2.0</td>
<td>2.2 3.6</td>
<td>460 1,400 81%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$2,760,000</td>
<td>0.97</td>
<td>2.1</td>
<td>2.4 4.4</td>
<td>850 1,490 79%</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>MMCW</td>
<td>$2,574,000</td>
<td>0.96</td>
<td>2.1</td>
<td>2.3 4.2</td>
<td>600 1,470 81%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$2,842,000</td>
<td>0.98</td>
<td>2.2</td>
<td>2.5 5.2</td>
<td>870 1,570 79%</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>MMCW</td>
<td>$2,720,000</td>
<td>0.98</td>
<td>2.2</td>
<td>2.5 4.8</td>
<td>730 1,560 80%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$2,943,000</td>
<td>0.99</td>
<td>2.3</td>
<td>2.6 6.2</td>
<td>960 1,680 80%</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>MMC</td>
<td>$2,956,000</td>
<td>0.47</td>
<td>1.8</td>
<td>1.0 1.7</td>
<td>60 1,300 92%</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>MMCW</td>
<td>$3,730,000</td>
<td>0.80</td>
<td>2.0</td>
<td>1.9 3.2</td>
<td>790 2,110 85%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$4,159,000</td>
<td>0.95</td>
<td>2.2</td>
<td>2.5 5.5</td>
<td>1,030 2,390 86%</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>MMCW</td>
<td>$4,125,000</td>
<td>0.92</td>
<td>2.1</td>
<td>2.4 4.3</td>
<td>980 2,300 78%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$4,429,000</td>
<td>0.98</td>
<td>2.3</td>
<td>2.6 6.4</td>
<td>1,250 2,540 82%</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>MMCW</td>
<td>$4,460,000</td>
<td>0.95</td>
<td>2.1</td>
<td>2.6 4.9</td>
<td>1,490 2,430 75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$4,632,000</td>
<td>0.99</td>
<td>2.4</td>
<td>3.0 7.4</td>
<td>1,270 2,730 80%</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>MMCW</td>
<td>$4,688,000</td>
<td>0.98</td>
<td>2.2</td>
<td>2.8 5.9</td>
<td>1,560 2,630 73%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MMCW-A</td>
<td>$4,930,000</td>
<td>0.99</td>
<td>2.4</td>
<td>3.1 8.5</td>
<td>1,330 2,970 75%</td>
</tr>
</tbody>
</table>

Table 4 Comparing different service levels for both MMCW and MMCW-A solved using 12-hr IPBLS.

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 load dispatch frequencies. Naturally, this also leads to higher on-time probabilities in solutions found using the MMCW-A model, as measured by the increase in vOTP. A retail shipper is likely to use two main metrics to assess a consolidation network design: (i) cost and (ii) service level. While \( p \) is a measure of worst-case service level performance of a design, vOTP is a measure of its average performance and may be a more relevant metric for customer service. In Table 4, we first observe that design solutions resulting from only minimizing cost would result in a vOTP service level of 47%. We then see a significant improvement in vOTP 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.

A retailer can use this type of analysis to decide which design solution best balances cost and customer service. Importantly, using the MMCW-A models directly to find network designs may
be a reasonable approach in some cases. For example, when assessing the Group 6 instances, the shipper may choose to use the MMCW-A solution with $p = 0.5$ because it is simple to set the conservatism levels (i.e., all route types use $c_r = 0.5$) and the cost is only an 11.5% increase for an 18.8% increase in service; note that this solution achieves a 3% higher vOTP than the MMCW solution with $p = 0.6$ at an insignificant cost increase. However, if they are dedicated to ensuring the best service, they may prefer to use the MMCW-A model with $p = 0.7$ solution, which has the lowest cost for an on-time probability of 99%.

5.4. Exact MIP versus Heuristic Solution Approach Performance

We now present results that verify that our single- and two-phase IP-based heuristic solution approaches are effective at solving realistically-sized problems. A best known lower bound is used to calculate MIP gaps, where this bound is computed 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 5 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 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 6. 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.

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
Table 5  Comparing MIP vs single-phase IPBLS heuristic performance.

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

(a) MMCW-A

<table>
<thead>
<tr>
<th>Group</th>
<th>2-week MIP Lower Bound</th>
<th>MIP 12-hr Obj</th>
<th>MIP Gap</th>
<th>IPBLS Heuristic 12-hr Obj</th>
<th>IPBLS Gap</th>
<th>Time to Obj (hr)</th>
<th>Heuristic Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$118,000</td>
<td>$118,000</td>
<td>0.0%</td>
<td>$118,000</td>
<td>0.0%</td>
<td>0.1</td>
<td>0.0%</td>
</tr>
<tr>
<td>2</td>
<td>$299,000</td>
<td>$311,000</td>
<td>3.9%</td>
<td>$311,000</td>
<td>4.0%</td>
<td>3.3</td>
<td>-0.1%</td>
</tr>
<tr>
<td>3</td>
<td>$625,000</td>
<td>$686,000</td>
<td>8.9%</td>
<td>$693,000</td>
<td>9.8%</td>
<td>9.5</td>
<td>-1.0%</td>
</tr>
<tr>
<td>4</td>
<td>$1,355,000</td>
<td>$1,536,000</td>
<td>11.7%</td>
<td>$1,522,000</td>
<td>11.0%</td>
<td>12.0</td>
<td>0.9%</td>
</tr>
<tr>
<td>5</td>
<td>$2,046,000</td>
<td>$2,359,000</td>
<td>13.2%</td>
<td>$2,273,000</td>
<td>10.0%</td>
<td>12.0</td>
<td>3.6%</td>
</tr>
<tr>
<td>6</td>
<td>$3,267,000</td>
<td>$4,226,000</td>
<td>22.7%</td>
<td>$3,730,000</td>
<td>12.4%</td>
<td>12.0</td>
<td>11.8%</td>
</tr>
<tr>
<td>7</td>
<td>$4,331,000</td>
<td>$6,046,000</td>
<td>28.4%</td>
<td>$5,096,000</td>
<td>15.0%</td>
<td>12.0</td>
<td>15.7%</td>
</tr>
<tr>
<td>8</td>
<td>$9,171,000</td>
<td>$19,633,000</td>
<td>53.3%</td>
<td>$12,512,000</td>
<td>26.7%</td>
<td>12.0</td>
<td>36.3%</td>
</tr>
<tr>
<td>9</td>
<td>$15,327,000</td>
<td>$39,720,000</td>
<td>61.4%</td>
<td>$25,969,000</td>
<td>41.0%</td>
<td>12.0</td>
<td>34.6%</td>
</tr>
</tbody>
</table>

(b) MMCW

Table 6  MMCW-A and MMCW solved using single-phase IPBLS compared to the two-phase IPBLS approach for solving MMCW.

<table>
<thead>
<tr>
<th>Group</th>
<th>MMCW MIP 2-week LB</th>
<th>MMCW-A MIP</th>
<th>MMCW MIP</th>
<th>8-hr MMCW-A + 4-hr MMCW MIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>$9,171,000</td>
<td>$12,457,000</td>
<td>26.4%</td>
<td>$12,512,000</td>
</tr>
<tr>
<td>9</td>
<td>$15,327,000</td>
<td>$22,911,000</td>
<td>33.1%</td>
<td>$25,969,000</td>
</tr>
</tbody>
</table>

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.

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 \( \frac{W_r}{|r|} \)); 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.
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.

A 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. Another extension is to 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. Finally, 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: 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 $R_k$, is depicted in Figure 7. Finally, Figure 8 illustrates the freight transportation costs introduced in Table 2 on a 500-mile leg.

<table>
<thead>
<tr>
<th>Origins</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>200 lbs</td>
<td>300 lbs</td>
<td>400 lbs</td>
</tr>
<tr>
<td>VND</td>
<td>0.02 TL</td>
<td>0.03 TL</td>
<td>0.03 TL</td>
</tr>
<tr>
<td>Medium</td>
<td>700 lbs</td>
<td>1,200 lbs</td>
<td>1,700 lbs</td>
</tr>
<tr>
<td>VND</td>
<td>0.06 TL</td>
<td>0.1 TL</td>
<td>0.15 TL</td>
</tr>
<tr>
<td>Large</td>
<td>1,400 lbs</td>
<td>2,300 lbs</td>
<td>3,700 lbs</td>
</tr>
<tr>
<td>VND</td>
<td>0.12 TL</td>
<td>0.19 TL</td>
<td>0.31 TL</td>
</tr>
<tr>
<td>FC</td>
<td>3,900 lbs</td>
<td>3,900 lbs</td>
<td>3,900 lbs</td>
</tr>
<tr>
<td></td>
<td>0.33 TL</td>
<td>0.33 TL</td>
<td>0.33 TL</td>
</tr>
</tbody>
</table>

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

Figure 7 Distribution of average allowable waiting delays for commodities across all instances.

Figure 8 Freight mode costs for a 500-mile leg.
References


