Middle Mile Consolidation Network Design with Robust Lead Time Constraints

Lacy Greening*, Mathieu Dahan, and Alan Erera
Georgia Institute of Technology, Atlanta, GA, USA
*Corresponding author, lacy.greening@gatech.edu

April 4, 2022

Keywords: e-commerce logistics; service network design; middle mile;

1 INTRODUCTION

New e-commerce retailing models such as ship-to-home and ship-to-store require that retailers implement new approaches for distributing directly to consumers on demand. Some high-demand stock-keeping units (SKUs) may be held at fulfillment centers (FCs), while others may be shipped directly from vendors. *Middle-mile consolidation network design* problems arise when large retailers jointly coordinate shipments from vendors and FCs into last-mile distribution (LMD) facilities, seeking consolidated truckloads when possible to reduce high less-than-truckload (LTL) and parcel freight charges. Competition among e-commerce retailers to serve consumers rapidly leads to tight lead time requirements for shipments in the middle mile. Traditional flat-network service network design (SND) models for trucking (Powell & Sheffi, 1983) can incorporate transit-time constraints, but do not account for delays created by low shipment frequencies. Time-expanded network models (Jarrah *et al.*, 2009, Erera *et al.*, 2013) can address this shortcoming, but often result in very large model instances that are intractable for the problems faced by larger shippers. In this work, we adopt an idea from public transit, i.e., model waiting delays as the inverse of outbound departure frequencies (Spiess & Florian, 1989, Cancela *et al.*, 2015), to fully capture constraints on shipment lead times in flat network models that can be solved effectively and demonstrate the impact of more conservative lead time constraints.

2 OPTIMIZATION MODELING

We consider a large shipper that needs to move shipments to fill orders from known *origins* to known *destinations* within specified lead times. Let $(\mathcal{N}, \mathcal{L})$ define the shipper's service network, where 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. The directed arc set $\mathcal{L}$ consists of the set of potential freight transportation lanes connecting pairs of locations, where a lane is defined as the physical arc with a transportation mode (e.g., LTL or truckload) assigned. Each load of size q dispatched on lane l incurs a fixed-plus-linear cost given by the expression $A_l + B_l q$. Furthermore, each lane specifies an associated upper bound Q_l^{max} and lower bound Q_l^{min} on the size of each dispatched load. Shipment demand is modeled using a set $\mathcal{K}$ of *commodities.* Let $\mathcal{R}_k$ represent the set of potential consolidation routes for commodity k, where each route is an ordering of adjacent freight transportation lanes connecting its origin o_k to its destination d_k, and potentially uses one or more transfer facilities. A unique freight route $r \in \mathcal{R}_k$ must be selected to specify a consolidation plan for commodity k and has a total handling cost of C_r. We denote $\mathcal{R} := \cup_{k \in \mathcal{K}} \mathcal{R}_k$ as the set of potential freight routes.

We use a flat network representation of capacity allocation to lanes and an associated representation of shipment consolidation into load dispatches. Thus, freight transportation capacity decisions are modeled as frequencies of load dispatches on lanes per time. The demand inputs are also expressed as constant rates per time; let V_k be the demand rate for commodity k, representing the aggregated average shipment size (volume) flowing from o_k to d_k per time. The goal 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. Thus, let binary variables x_r indicate whether route $r \in \mathcal{R}$ is selected, continuous variables v_l indicate the total shipment volume assigned to each lane l, and integer variables f_l count the number of loads dispatched per time on lane l. We can formulate the *middle mile consolidation* (MMC) model as follows:

$$\min_{x,f,v} \quad \sum_{r\in\mathcal{R}} C_r x_r + \sum_{l\in\mathcal{L}} \big(A_l f_l + B_l v_l\big) \tag{1a}$$

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

$$v_l = \sum_{k\in\mathcal{K}} \sum_{\{r\in\mathcal{R}_k | r\ni l\}} V_r x_r, \qquad \forall\, l \in \mathcal{L}, \tag{1c}$$

$$v_l \le Q_l^{max} f_l, \qquad \forall\, l \in \mathcal{L}, \tag{1d}$$

$$v_l \ge Q_l^{min} f_l, \qquad \forall\, l \in \mathcal{L}. \tag{1e}$$

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 lane l aggregated across commodities. Constraints (1d) and (1e) set the required load dispatch frequencies for each lane using upper and lower bounds on load size.

2.1 Using Waiting Delays to Constrain Lead Times

The frequency of load dispatches on a lane impacts lead times since lower frequencies lead to longer waiting delays at dispatch locations. Given the load dispatch frequency f_l on lane l, we assume trucks are scheduled to deterministically dispatch every $\frac{1}{f_l}$ time units and are uncoordinated across facilities. If the shipments for each commodity k are then assumed to arrive ready for shipment at o_k according to a uniform distribution, it is reasonable to assume that all arriving shipments to be dispatched from a facility will have similarly distributed arrival times. Thus, the waiting delay experienced by any individual shipment on each lane l can be modeled as a uniform random variable $W_l \sim \text{Uniform}(0, \frac{1}{f_l})$. Given this model, the expected waiting time experienced by all shipments before being dispatched on lane l is $\mathbb{E}[W_l] = \frac{1}{2}\frac{1}{f_l}$. We define the expected lead time of a route as the sum of lane transit times and expected waiting delays for load dispatches. The allowable waiting delay $\hat{W}_r$ of route r is its lead time requirement less the sum of its lane transit times. A load plan satisfies the lead time requirement (in expectation) of route r if and only if the total expected waiting delay along that route does not exceed $\hat{W}_r$ (i.e., $\sum_{l\in r} \frac{1}{2}\frac{1}{f_l} \le \hat{W}_r$).

To formulate this *middle mile consolidation with waiting times* (MMCW) model, we use a linearization approach similar to that proposed in Cancela *et al.* (2015) for transit network design problems. Define for each lane l a set of binary variables z_{lf} over a finite set $\mathcal{F}_l$ of possible load dispatch frequency values. We then substitute the frequency variables as follows:

$$f_l = \sum_{f\in\mathcal{F}_l} f z_{lf}, \quad \forall\, l \in \mathcal{L}. \tag{2}$$

We now formulate the MMCW model as follows:

$$\begin{aligned}
\min_{x,z,v} \quad & \sum_{r\in\mathcal{R}} C_r x_r + \sum_{l\in\mathcal{L}} \big[A_l\big(\sum_{f\in\mathcal{F}_l} f z_{lf}\big) + B_l v_l\big] && \text{(3a)}\\
\text{s.t.} \quad & \text{(1b)} - \text{(1e) using (2) where applicable,} && \\
& \sum_{f\in\mathcal{F}_l} z_{lf} \le 1, \qquad \forall\, l \in \mathcal{L}, && \text{(3b)}\\
& \frac{1}{2}\sum_{l\in r}\sum_{f\in\mathcal{F}_l} \frac{1}{f} z_{lf} \le \hat{W}_r x_r + \frac{|r|}{2}(1-x_r), \qquad \forall\, r\in\mathcal{R}. && \text{(3c)}
\end{aligned}$$

Constraints (3b) select at most one frequency for each lane. Constraints (3c) ensure the total expected waiting time along each selected route is no more than the total allowable waiting delay. Note that the main drawback of the MMCW model is the potentially large number of binary variables z_{lf} needed when lanes have many possible frequency values. To avoid this, we now develop a simpler alternative for finding a reasonable solution to the MMCW problem denoted as the *middle mile consolidation with allocated waiting delay* (MMCW-A) model. This approach restricts the space of feasible solutions by allocating fixed fractions of a route's total allowable waiting delay *a priori* to each of its lanes. Under this assumption, the following linear constraints can be directly added to (1) to yield the MMCW-A model:

$$f_l \ge \frac{1}{2}\frac{|r|}{\hat{W}_r} x_r, \quad \forall\, r\in\mathcal{R},\ \forall\, l\in r. \tag{4}$$

The right-hand side of constraints (4) represents the minimum number of load dispatches needed to ensure the expected waiting delay for each load dispatched does not exceed the evenly-distributed allowable waiting delay. The simpler MMCW-A model can be used to find a strong starting solution for the MMCW model as the commodity and/or route sets grow much larger.

2.2 Protecting Against Variability in Waiting Delays

Constructing the lead time constraints (3c) or (4) using the expected waiting delays $\mathbb{E}[W_l] = \frac{1}{2}\frac{1}{f_l}$ does not account for the potential variance in the random load dispatch waiting delays and can only guarantee that the lead time requirements are met with probability 0.5. However, for every route r and desired commodity on-time arrival probability p_r, the shipper may want to select load dispatch frequencies f_l for each lane $l \in r$ so that $\mathbb{P}(\sum_{l\in r} W_l \le \hat{W}_r) \ge p_r$, that is, the probability of the total waiting delay on route r not exceeding $\hat{W}_r$ is at least p_r. Since the waiting delay experienced by arriving shipments for lane 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):

$$\mathbb{P}\left(\sum_{l\in r} W_l \le \hat{W}_r\right) = \frac{1}{|r|!\prod_{l\in r}\frac{1}{f_l}} \sum_{J\subseteq r} (-1)^{|J|}\left[\max\left\{0, \hat{W}_r - \sum_{l\in J}\frac{1}{f_l}\right\}\right]^{|r|}. \tag{5}$$

Expression (5) is nonlinear in the load dispatch frequencies and again cannot be included directly in the optimization model (1). To circumvent this limitation, we define for each route $r \in \mathcal{R}$ a conservatism hyperparameter $c_r \in [0, 1]$ that the shipper can vary to adjust the probabilistic guarantee of meeting the commodity lead time requirement. Then, for every route $r \in \mathcal{R}$, we respectively replace constraints (3c) and (4) by

$$c_r \sum_{l\in r}\sum_{f\in\mathcal{F}_l} \frac{1}{f} z_{lf} \le \hat{W}_r x_r + c_r |r|(1-x_r), \tag{6}$$

$$\text{and}\quad f_l \ge c_r \frac{|r|}{\hat{W}_r} x_r, \quad \forall\, l\in r. \tag{7}$$

Thus, we determine the lowest hyperparameter for which the lead time constraints (6) or (7) remove any load plan that does not meet the desired probability p_r. Interestingly, we observe that deriving the conservatism level c_r for the MMCW-A approach simply requires the on-time probability p_r and number of legs $|r|$, and is independent of the allowed waiting delay $\hat{W}_r$.

3 RESULTS

The instances used are synthetic but have been derived from historical weekly demand data provided by a large U.S.-based e-commerce retailer. The instances have 160 vendors, 8 FCs, and 90 LMD facilities with over 10,000 commodities; each commodity can have up to 5 route options. Given the computational difficulty of solving a problem of this size, we run an IP-based local search (IPBLS) for 12 hours to find good solutions. To measure service level, we calculate the volume-weighted expected on-time probability (vOTP) of a solution as $\frac{\sum_{k\in\mathcal{K}}\mathbb{P}(\sum_{l\in r} W_l \le \hat{W}_r) V_k}{\sum_{k\in\mathcal{K}} V_k}$, where the on-time probability of an individual commodity using route r is calculated using (5), the assigned load dispatch frequencies $f_l \; \forall \, l \in r$, and the allowable waiting delay $\hat{W}_r$. Table 1 shows the results of using four different values of on-time probability guarantees.

Min p_r	Model	12-hr IPBLS Obj	vOTP	Vol-Wtd Route Length	Avg Load Disp Freq (#/week)		Loads/Week		Vol-Wtd Utilization
					LTL	TL	LTL	TL	TL
0	MMC	\$2,956,000	0.47	1.8	1.0	1.7	60	1,300	92%
0.5	MMCW	\$3,730,000	0.80	2.0	1.9	3.2	790	2,110	85%
	MMCW-A	\$4,159,000	0.95	2.2	2.5	5.5	1,030	2,390	86%
0.6	MMCW	\$4,125,000	0.92	2.1	2.4	4.3	980	2,300	78%
	MMCW-A	\$4,429,000	0.98	2.3	2.6	6.4	1,250	2,540	82%
0.7	MMCW	\$4,460,000	0.95	2.1	2.6	4.9	1,490	2,430	75%
	MMCW-A	\$4,632,000	0.99	2.4	3.0	7.4	1,270	2,730	80%
0.8	MMCW	\$4,688,000	0.98	2.2	2.8	5.9	1,560	2,630	73%
	MMCW-A	\$4,930,000	0.99	2.4	3.1	8.5	1,330	2,970	75%

Table 1 – *Comparing different service levels for MMCW and MMCW-A solved using 12-hr IPBLS.*

We first observe that a design solution resulting from minimizing cost only 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. We also note that the number of loads dispatched increases as p_r increases, which results in less utilization of the trucks sent. A retailer can use this type of analysis to decide which design solution best balances cost and customer service. Finally, using the restricted MMCW-A models to find network designs is a reasonable approach in some cases.

References

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

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

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

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

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

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