

## INFORMS Journal on Applied Analytics

Publication details, including instructions for authors and subscription information:  
<http://pubsonline.informs.org>

### Empowering MetroAccess Service with Nested Decomposition and Service Type Integration

Shijie Chen; , Md Hishamur Rahman; , Nikola Marković; , Muhammad Imran Younus Siddiqui, Matthew Mohebbi, Yanshuo Sun;

To cite this article:

Shijie Chen; , Md Hishamur Rahman; , Nikola Marković; , Muhammad Imran Younus Siddiqui, Matthew Mohebbi, Yanshuo Sun: (2025) Empowering MetroAccess Service with Nested Decomposition and Service Type Integration. INFORMS Journal on Applied Analytics 55(3):238-253. <https://doi.org/10.1287/inte.2024.0123>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact [permissions@informs.org](mailto:permissions@informs.org).

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2024, INFORMS

Please scroll down for article—it is on subsequent pages





With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes. For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

# Empowering MetroAccess Service with Nested Decomposition and Service Type Integration

Shijie Chen,<sup>a</sup> Md Hishamur Rahman,<sup>b</sup> Nikola Marković,<sup>b</sup> Muhammad Imran Younus Siddiqui,<sup>c</sup> Matthew Mohebbi,<sup>c</sup> Yanshuo Sun<sup>a,\*</sup>

<sup>a</sup>Department of Industrial and Manufacturing Engineering, FAMU-FSU College of Engineering, Florida State University, Tallahassee, Florida 32310; <sup>b</sup>Department of Civil and Environmental Engineering, University of Utah, Salt Lake City, Utah 84112; <sup>c</sup>IT Curves, Gaithersburg, Maryland 20879

\*Corresponding author

Contact: sc20hw@fsu.edu,  <https://orcid.org/0000-0003-4396-1357> (SC); hisham.rahman@utah.edu (MHR); nikola.markovic@utah.edu (NM); myyounus@itcurves.net (MIYS); mmohebbi@itcurves.net (MM); y.sun@eng.famu.fsu.edu,  <https://orcid.org/0000-0003-2943-4323> (YS)

Received: April 14, 2024

Revised: August 4, 2024

Accepted: August 30, 2024

Published Online in Articles in Advance:  
October 24, 2024

<https://doi.org/10.1287/inte.2024.0123>

Copyright: © 2024 INFORMS

**Abstract.** Tens of millions of Americans face severe mobility barriers because of travel-limiting disabilities and thus depend on paratransit, a door-to-door shared-ride service. However, this crucial service faces significant operational and financial challenges. Therefore, through a university-industry collaboration, we have identified (1) modernizing the optimization engine in paratransit scheduling software suites, and (2) incorporating alternative service providers in paratransit service optimization as two critical steps in overcoming some challenges in paratransit practice. We thus developed a nested decomposition method in which a column generation-based solution approach is embedded in a temporal decomposition framework. Additionally, we integrated alternative service types, such as accessible taxis, in paratransit scheduling and designed a reoptimization procedure. The new optimization methods were implemented in a software suite of IT Curves and deployed in paratransit operations in the Washington, DC, metro area. It was found that the improved optimization engine, relative to the legacy, led to significant improvements in key operational metrics and yielded substantial operating cost reductions (approximately 15%). The fruitful collaboration not only showcases the potential for advanced algorithms to significantly enhance the financial sustainability of paratransit services but also reflects the importance of bridging the gap between theoretical research and practical application in transit vehicle routing.

**History:** This paper was refereed.

**Funding:** This work was supported by the National Science Foundation [Grant 2055347].

**Keywords:** mobility equity • individuals with disabilities • paratransit service • decomposition-based optimization • case study

## Introduction

As mandated by the Americans with Disabilities Act (ADA), every public transit operator in the United States must provide paratransit services for passengers who are unable to use conventional public transit because of their disabilities (Chen et al. 2024). ADA paratransit provides critical access for individuals with disabilities to essential services and resources such as employment, healthcare, education, and social engagement. The provision of paratransit greatly reduces the mobility disparity between the general public and individuals with disabilities and directly contributes to improved independence and community engagement for the otherwise isolated population group.

Despite its significance and benefits, ADA paratransit faces substantial financial challenges. As a specialized form of public transportation, ADA paratransit is substantially more expensive to operate than fixed-route transit services. A national survey on the transit

industry by the American Public Transportation Association (APTA) indicated that paratransit accounted for approximately only 2% of all public transportation trips in the United States while incurring approximately 9% of total capital and operating expenses (CBO 2024). The disproportionately high expense of paratransit was also reported in several major metropolitan areas in the United States. For instance, in the Washington, DC, metro area in 2022, paratransit costs were 11 times higher than those of bus services (WMATA 2022b).

Although paratransit operating costs remain high across the country, ADA-mandated fare regulations result in limited revenue, posing a challenge to the financial sustainability of these services. In Los Angeles, paratransit fare revenues account for just 4% of the overall operating budget in 2022 (Chicago Metropolitan Agency for Planning 2023), leading to a 96% financial deficit to be filled by public funds or subsidies. In 2022, the revenue of MetroAccess, the paratransit service in

the DC area, generated from a one-way fare constituted merely 2% of its operating budget (WMATA 2021). The low farebox recovery ratio was observed not only in major metropolitan areas but also at the national level. The average farebox recovery of paratransit service in 25 major agencies across the United States in 2019 was 7% (Shrode 2022).

Despite significant financial subsidies for paratransit, its customers often experience less-than-ideal service levels. One major issue is the service's limited responsiveness to demand, indicating an accessibility gap compared with other transportation options. Nationwide, paratransit riders typically must book their trips one day in advance, and their same-day trip requests or changes are rarely accommodated. By contrast, other on-demand mobility services (such as Uber) provide access within minutes; traditional transit modes (such as bus and rail) allow for on-the-spot travel during operational hours without the need for a reservation. Furthermore, paratransit operators often use wide scheduling windows, such as 90 minutes, which means that a rider could be picked up 45 minutes before or after their requested time. Once scheduled, riders are allocated a 30-minute window for pickup, within which they must be ready. The relatively uncertain pickup time gives the rider no control over the expected arrival time, which further means unnecessary waiting at the destination (when arriving too early) and missing appointments (when arriving too late). Finally, it is inconvenient for a rider to modify the trip, as substantial advance notice (such as one day prior) is required. The lack of responsiveness, reliability, and flexibility contributes to the overall poor experience of paratransit riders.

A multitude of reasons cause the heavy financial burden on the government and the poor travel experience of customers. First, ADA paratransit is demand responsive and door to door, which inherently demands more resources in terms of vehicles and driver time and therefore a higher per-trip cost. By contrast, fixed-route transit, such as bus service, benefits from comparatively higher demand density and thus can exploit economies of scale. Second, the need for ADA-compliant vehicles with accessibility features increases capital and maintenance costs. Third, the distinctive features of ADA paratransit require specialized vehicle scheduling and routing algorithms. Although sophisticated exact optimization methods for paratransit have been extensively documented in academic literature (Ho et al. 2018), their application in the paratransit practice is rarely found. The existing software suites adopted by paratransit operators mostly rely on fast heuristics, which, unfortunately, result in suboptimal paratransit service plans and thus high operating costs.

To elevate the travel experience of transportation-disadvantaged individuals and address the financial challenges faced by ADA paratransit operators, a team

consisting of researchers from Florida State University and the University of Utah has developed an innovative solution to the paratransit scheduling problem with research funding from the National Science Foundation (NSF) that aims to push the boundaries of paratransit scheduling software. Although paratransit suffers from other issues, this paper focuses on advancing paratransit vehicle routing and scheduling practice. The research contribution lies in two aspects: (1) designing more effective and scalable paratransit optimization algorithms, and (2) incorporating other service types in paratransit optimization, to be elaborated in detail next.

The primary motivation for this NSF project arises from the significant gap between research and practice in paratransit scheduling. Various advanced optimization techniques, such as branch and price and cut (Luo et al. 2019), have recently been proposed by academics to solve the paratransit scheduling problem exactly. However, commercial software suites for paratransit scheduling, such as the one developed by IT Curves, are largely based on decades-old heuristics mainly because of the ease of implementation and computational efficiency. The sophisticated algorithms developed in academia are often considered computationally prohibitive. For instance, Luo et al. (2019) solved an instance involving 36 rider requests in four hours. The largest instance solvable with the exact algorithm in Rist and Forbes (2021) has only 144 trips, whereas a typical instance faced by a major operator in the real world could consist of thousands of trips. To bridge this gap, university researchers have developed a new decomposition-based routing and scheduling algorithm, which strikes a desirable balance between achievable solution quality and needed computation effort.

Another motivation stems from the lack of a systematic and computerized method for identifying and assigning certain trip requests to additional mobility options, such as taxis and Uber, despite the increasing trend among paratransit operators to include these alternative options. For example, the transit authority serving the nation's capital area is reported to select trips randomly (Uber 2023), which is far from ideal. It should be noted that trip requests with similar spatio-temporal characteristics facilitate ridesharing and thus should not be assigned to other nonshared mobility options. Conversely, requests that “do not fit well” with the rest of the demand may imply significant (per-trip) cost because of extended operating hours in the case of temporal outliers and/or significant deadheading in the case of spatial outliers. Therefore, such outliers should be assigned to alternative mobility options instead (Rahman et al. 2023). There is no known documentation that any commercial paratransit scheduling software has incorporated this key functionality.

In partnership with IT Curves, the university team has implemented their next-generation algorithm based

on a nested decomposition approach and a reoptimization heuristic in Mobile Resource Management System (MRMS), a software suite for paratransit scheduling developed by IT Curves. Therefore, the current paratransit optimization engine in MRMS has been upgraded in place of the previous insertion heuristics, and alternative service options have also been incorporated under the same optimization framework, more specifically through the reoptimization procedure. Real-world deployments by Challenger Transportation (CT), which provides paratransit services in the DC area, have indicated that implementing the nested decomposition algorithm and enabling service type integration could jointly reduce the total service delivery cost by at least 15%. At the same time, substantial improvements are achieved in many key operational metrics, such as vehicle idling time.

The convincing improvements in several operational metrics and cost-effectiveness achieved in the Washington, DC, metro area suggest that similar benefits could be realized across the United States when the developed optimization approach is deployed in other metro areas. The successful academia-industry collaboration among Florida State University, the University of Utah, and IT Curves can also be adapted to address different emerging mobility challenges in other regions. Therefore, the operational efficiency and fiscal viability of ADA paratransit are expected to improve not only in the DC area but also across the United States, directly benefiting millions of transportation-disadvantaged riders whose lives depend on ADA paratransit.

## Practices of WMATA MetroAccess Service Practice

Established in 1967, the Washington Metropolitan Area Transit Authority (WMATA) was tasked with the planning, development, and management of public transit within the Washington, DC metropolitan region. In 1993, WMATA started MetroAccess, an ADA complementary paratransit service, covering the DC area and neighboring counties in Maryland and Virginia, including the independent cities of Alexandria and Falls Church (WMATA 2022a). By regulation, MetroAccess must be available within 0.75 miles of fixed transit (see the highlighted area in Figure 1) and operate during the same hours as fixed-route service. Although Loudoun County, Virginia, is not shown in Figure 1, MetroAccess is indeed available within 0.75 miles of three Silver Line Metrorail stations, which were opened in 2022. The five red points indicate the subcontractors' depots. The monthly ridership ranged from 110,000 to 130,000 as of 2023 in the nation's sixth-largest metro area.

Eligible customers of MetroAccess, determined by their inability to use fixed-route public transit, can request trips one to seven days in advance through an online booking system or by phone. However, they

must submit their requests no later than 4:30 p.m. the day before the trip and do not have the option for same-day service requests. Trips are scheduled within a predetermined window around the requested pickup time, specifically 45 minutes before through 45 minutes after. Vehicles must arrive within the pickup time window to be considered on time, which is a key performance metric adopted by MetroAccess. The service also has strict policies about rider no-shows and late cancellations, which may lead to penalties such as service suspension (WMATA 2023a).

## Abilities-Ride Program (ARP)

WMATA also offers the Abilities-Ride Program as a flexible alternative, allowing MetroAccess customers to use local taxis or Uber, among other mobility services. Although the ARP service is curb to curb, not door to door as in MetroAccess, ARP riders may enjoy a fast and direct trip to their destinations, not sharing the ride with other passengers. Although a customer may elect a preferred ARP provider, it is up to WMATA to determine whether a request will remain in MetroAccess or get assigned to an ARP provider one day before the service. The ARP fare policy, last updated in 2023, sets no additional fare for ARP (WMATA 2023a).

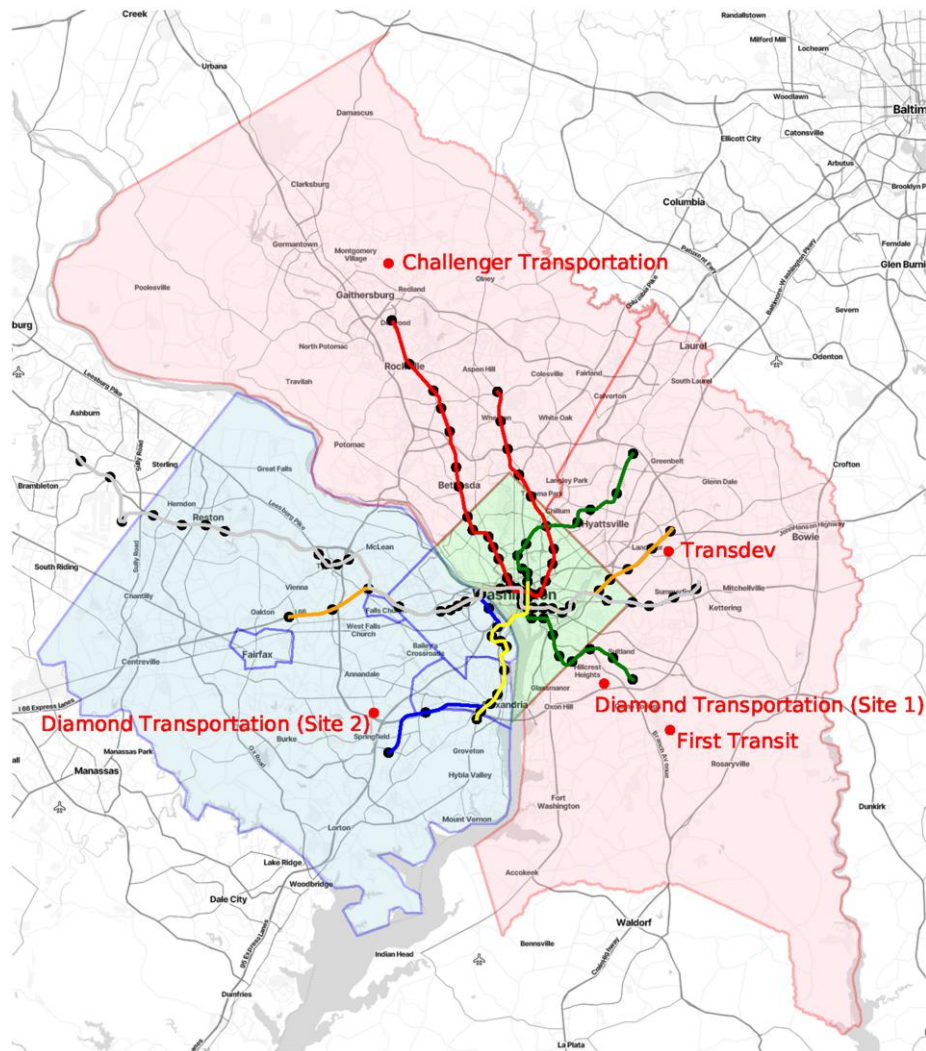
Although WMATA does not disclose how trips are split between MetroAccess and Abilities-Ride, a provider of the ARP, namely Uber, reports that the trip selection process is random, although some criteria, such as pickup times and trip lengths, are considered (Uber 2023). WMATA selects approximately 10%–15% of trips nightly for alternative providers such as Uber, which accounts for around 800 trips (Uber 2023). The random selection process may sound fair to service providers; nonetheless, it can lead to suboptimal operating efficiency. Hence, the practice can be significantly advanced with a systematic approach should it be successfully integrated into the paratransit scheduling and routing software.

## Service Delivery Model

MetroAccess has undergone significant transformations in its service delivery model over the past decades to accommodate demand growth and enhance service quality. The service was initially delivered through contracts with a single service broker called LogistiCare, Inc. from 2000 to 2005, which was replaced by MV Transportation, Inc. (MV) in 2006. MV was responsible for 50% of MetroAccess operations, including managing a call center to accept trip reservation requests, organizing and allocating these requests to carriers, coordinating vehicle dispatches, and overseeing field operations (Planners Collaborative, Inc. 2007). The remainder of the service was subcontracted by MV to other service carriers.



**Figure 1.** (Color online) MetroAccess Is Available Mainly Within 0.75 Miles of Fixed-Route Transit (Rail and Bus) in the Colored Counties/Cities as of May 2023



*Note.* Only Metrorail lines are shown; Metrobus routes are omitted because of their extensiveness.

In 2013, WMATA introduced a pivotal change in its paratransit service delivery model (Transit Cooperative Research Program 2018). The strategic shift was driven by the need to improve service efficiency, safety, and customer satisfaction in response to the doubled ridership of MetroAccess and the anticipated continuous demand growth because of the increasingly aging population and a growing number of people with disabilities. The new model transitioned from a single-contractor system to a more diversified and competitive framework involving multiple contractors. The three dedicated service providers were Transdev, First Transit, and Diamond (Transit Cooperative Research Program 2018), which maintained a 50%/35%/15% service split. The transition aimed to leverage the specialized capabilities of various service providers, fostering innovation and flexibility in service delivery while maintaining high

standards for safety and on-time performance. Contractors were subject to higher performance standards, with greater disincentives for failing to meet these criteria, such as the on-time performance threshold set at 92%. This model also allowed WMATA to address individual contractor performance issues more effectively, ensuring that only those who met or exceeded standards could expand their service scopes.

Challenger Transportation, established in 2000, is a dedicated service provider that has adeptly thrived under WMATA's evolving service delivery models. Initially a subcontractor to LogistiCare in 2000 and later to MV Transportation in 2006, CT demonstrated its adaptability and expanded its fleet and operations significantly. In 2018, CT was selected as a direct subcontractor to MetroAccess, responsible for 15% of the total load, subsequently increasing its service delivery to

**Figure 2.** (Color online) Some Vendors of Paratransit Scheduling Software Suites



more than 23%. The milestone marked a significant achievement for CT. As of 2021, the majority of MetroAccess trips were provided in Montgomery and Prince George’s counties in Maryland (WMATA 2021). CT’s depot location in Montgomery County, Maryland, as shown in Figure 1, establishes itself as a key subcontractor of MetroAccess.

**Optimization Software Suites**

All major paratransit operators, including subcontractors of MetroAccess such as CT, use computer-aided paratransit scheduling and dispatch software for their service delivery (Kessler 2004). Some software suites are presented in Figure 2, based on a thorough examination of the marketing materials and exhibits of paratransit scheduling software vendors in the 2023 TRANSform Conference & EXPO that was organized by the APTA.

Among them, the MRMS is a software suite available to CT that is developed and maintained by IT Curves.

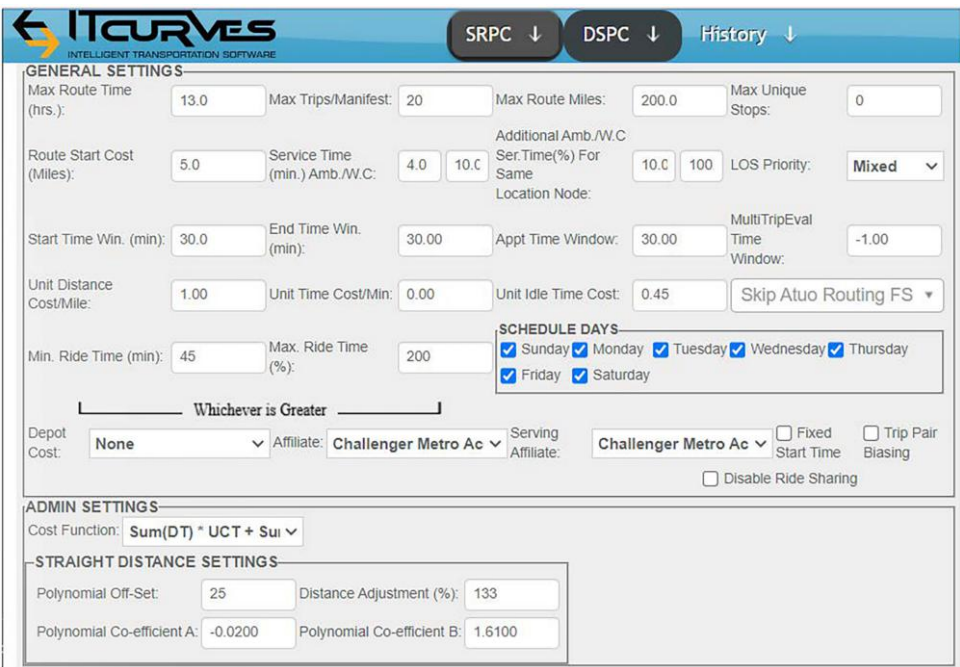
Established in 2008, IT Curves specializes in transportation automation technologies, including ADA paratransit scheduling, smartphone booking apps, and real-time tracking services. MRMS represents a typical technological advancement in the field of paratransit management, as it incorporates various algorithms to optimize the scheduling and routing of vehicles, aiming to enhance operational efficiency, customer service, and responsiveness to dynamic customer needs (Mohebbi et al. 2023). MRMS is used by dozens of other paratransit operators such as CT throughout the United States.

Figure 3 shows the MRMS interface for parameter settings. The paratransit scheduling engine employed within MRMS is mainly based on an insertion-based heuristic (to be described in a Section “Parallel Insertion Heuristic”), modified to incorporate operational requirements such as blocks, locked blocks, and outlier dispatching (Marković et al. 2015). Figure 4 shows an example screenshot of the MRMS.

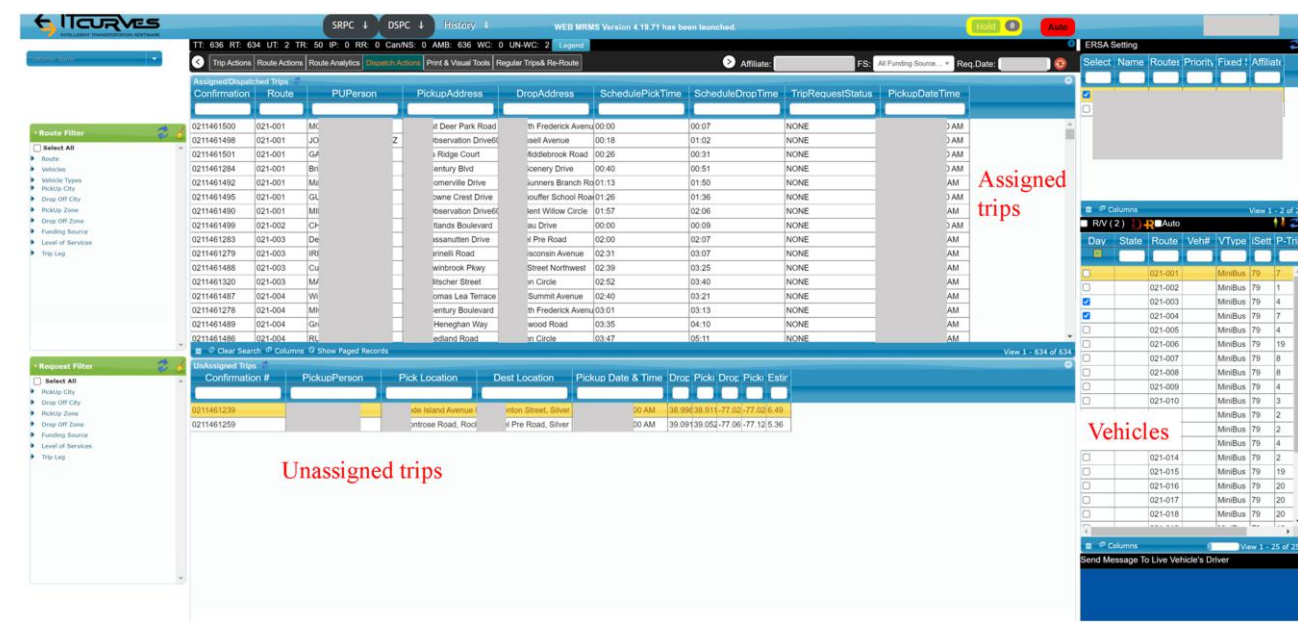
**ADA Paratransit Routing and Scheduling Problem**

The ADA Paratransit Routing and Scheduling Problem is commonly referred to as the Dial-A-Ride Problem (DARP), which is considered a variant of the pickup and delivery problem (PDP) in the optimization literature (Hanne et al. 2009). We follow Ho et al. (2018) and present a high-level definition of the static DARP as follows. The paratransit operator has a set of trip requests; each of which is characterized by pickup and drop-off locations, a requested time window for either pickup or drop-off, and the number of passengers (i.e., party size).

**Figure 3.** (Color online) The Parameters for Paratransit Scheduling Are Specified in This Control Panel of the MRMS



**Figure 4.** (Color online) This Screen Shows Not All Trip Requests Can Be Routed Given the Fleet Size and Vehicle Capacity Constraints



Notes. Two requests remain unassigned. Illustrative data are partially shown to protect data privacy.

The operator employs a fleet of vehicles located at a common depot to pick up and drop off paratransit riders. Typical constraints involved in the picking-up and delivering decisions are precedence (pickup prior to drop-off), time window, vehicle capacity, route duration, and maximum ride time constraints. A typical optimization objective is to minimize the total vehicle routing cost or time.

The DARP is distinct from other PDPs mainly because of its human perspective. Paratransit riders have specific preferences and special needs that must be addressed by the operator as regulated by the ADA. That explains why the maximum ride time constraint is especially relevant in the DARP. Conversely, in other PDPs, items being transported usually have no preferences and are not subject to similar regulations.

In a conventional DARP, only shared paratransit vehicles are employed for operations. Emerging studies have further considered alternative mobility options, such as Uber or taxis (Rahman et al. 2023). The integration of multiple fleet types for paratransit is intended to efficiently accommodate certain trip requests that are either spatial or temporal outliers that do not fit well with the rest of paratransit trip requests. The provision of alternative service providers thus contributes to improved operational efficiency, albeit increasing the complexity of vehicle routing and scheduling.

### Parallel Insertion Heuristic

In the 1980s, Jaw et al. (1986) introduced the first insertion-based algorithm for the DARP, called the

sequential insertion heuristic (SIH), which was computationally efficient, although its solution quality was not satisfactory. This heuristic was later upgraded by Toth and Vigo (1996) to the widely known algorithm called the parallel insertion heuristic (PIH). The core shared by both heuristics is to find the “cheapest” insertion position for a trip under consideration, whereas they differ in which trip is up for insertion. At the beginning of both heuristics, all trip requests are sorted by the requested pickup time, and the earliest request is used to initialize the first route. In the SIH, the next unscheduled trip is inserted individually into one of the available routes that yields the smallest cost increase. In the PIH, however, multiple trips in a look-ahead window of fixed length are considered simultaneously for insertion, although only one trip with the least cost increase is selected for insertion. After an insertion, the set of trips in the look-ahead window is updated for the next insertion. In both heuristics, all given routes are readily available to accommodate trip requests, including those without any assigned riders. By contrast, in MRMS’s implementation of the PIH, a new route is activated only when there exist no feasible insertions among activated routes, mainly to avoid the significant fixed cost of building a new route (also called startup cost in MRMS). The heuristics are terminated when all trips have been inserted (i.e., routed and scheduled). As the PIH yields better solutions than the SIH, this heuristic is preferred and widely adopted in practice, including by IT Curves in its MRMS.



Despite its selection for implementation in MRMS, the PIH has a few inherent shortcomings. First, it cannot revise previous routing decisions based on new information that becomes available after the insertion and thus lacks a corrective mechanism. Second, it misses the opportunity to create more efficient route segments by simultaneously inserting multiple trips into the windows of one existing route. Third, the PIH requires significantly more time than the SIH, especially when the length of the moving window is substantial. The PIH may run for hours to solve a realistic DARP consisting of close to 1,000 trips. Finally, the current PIH is unable to consider alternative mobility options that are increasingly preferred by riders and can reduce the total service delivery cost. Therefore, there is a pressing need to modernize the vehicle scheduling and routing engine in MRMS.

### Next-Generation Paratransit Optimization Method

The new optimization engine in the MRMS is powered by a sophisticated operations research framework consisting of a nested decomposition approach and a reoptimization procedure. An initial solution to the static paratransit optimization problem is obtained via well-designed decomposition and further improved by a heuristic that considers service type integration. Table A.1 in the appendix contains a list of the symbols and abbreviations used in the paper.

#### Nested Decomposition Approach

The decomposition approach is a two-level process (hence named “nested”) for partitioning a large-scale optimization problem into smaller and more manageable components, each corresponding to a single time period. First, a rolling-horizon method is used to achieve temporal decomposition. Second, column generation is employed to partition a problem associated with a given time period into a master problem and a subproblem, thus avoiding enumerating an exponential number of possible solutions.

**Temporal Decomposition.** Decomposing a large problem by time is naturally appealing. In the case of paratransit scheduling, early morning trip requests could be safely separated from late afternoon requests because of their substantial time difference. The separation is valid because these requests from vastly different time periods are unlikely to interact with each other, thereby warranting independent consideration. Nonetheless, when the time difference is less significant, the independence is undermined, and some trips may become interrelated, such as late morning and early afternoon trips. While dividing a single day into a series of periods, it is necessary to consider the interactions among trips in

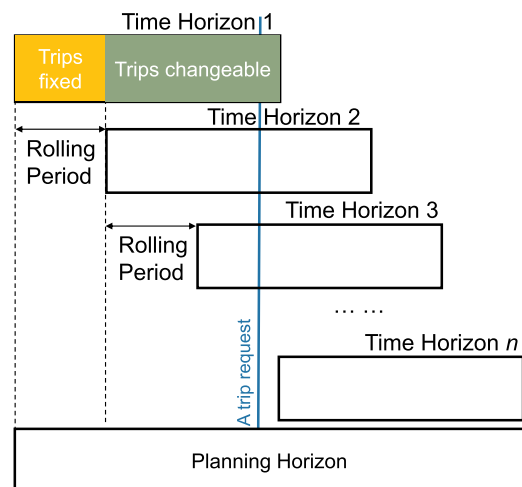
adjacent time periods to mitigate the compromise of solution quality because of temporal decomposition.

The trip interactions are captured by constructing reasonable overlap between adjacent time periods, which enables the decisions made for some trips in one period to be possibly corrected in the next period when later trip requests become available. In other words, the overlap ensures when the optimization process concludes in one period, some requests could be postponed to the next time period, and decisions regarding those postponed requests could be reevaluated and readjusted. However, because of a limit on the number of times an order gets postponed, the impact of postponed requests on the subsequent epochs is diminishing, which mirrors the notion of *slow cooling* in the simulated annealing metaheuristic, where the exploration of solution space gradually shrinks as the temperature progressively drops over time.

We then describe the rolling horizon approach as follows. An entire planning horizon is decomposed into a series of time horizons (THs) as shown in Figure 5. Except for those few trips at the very beginning or end of the planning horizon, a trip request falls into multiple THs when the rolling period (RP) is shorter than a TH. For each TH, a DARP can be formulated and solved involving all trips in this TH. However, depending on whether decisions regarding a trip must be fixed right now or can be revisited in later THs, there are two types of trips. When a trip does not fall into the immediate next TH, its relevant decisions must be finalized; otherwise, the preliminary decisions made in the current TH could be reoptimized, considering other trips in the next TH.

Two key parameters in the previous temporal decomposition, namely, RP and TH, determine the delicate balance between computational efficiency and solution

**Figure 5.** (Color online) A Planning Horizon Can Be Decomposed into Multiple Time Horizons with Carefully Selected Overlaps





quality. When RP is fixed while TH increases, trips in a larger time span are simultaneously considered for solving a DARP, thus yielding a better solution quality but requiring more computation time. In particular, when TH grows to be the same as the whole planning horizon, the benefit of decomposition reduces to none. If one fixes TH while increasing RP, the degree of overlap decreases, thus reducing the potential to find better solutions because of less corrective optimization. However, increasing RP reduces the total computation time. In the extreme case in which RP equals TH, no trips would be revisited, and suboptimal solutions are inevitable, although the total solution time is minimized because of the elimination of solution revisitation. In general, the RP could range from 30 to 60 minutes. Preferably, the TH should be set at four to six times the average trip length (approximately 15 minutes per trip in our study). When the step size for parameter searching is further specified as 10 or 15 minutes, there are only a limited number of possible settings for TH and RP. A simple grid search is thus recommended for parameter tuning.

Although the rolling horizon approach has been extensively used in solving dynamic problems such as time slot management (Bruck et al. 2020) and container transportation scheduling problems (Zhen et al. 2022), its application to the solution of a large-scale static optimization problem is novel and proven to be highly effective in practice.

**Set Partitioning Problem.** For each time horizon  $t$ , we seek to construct an optimization problem, known as the set partitioning problem. Until TH  $t$ , a set of vehicle routing and scheduling plans (simply plans)  $\bar{\Omega}_t$  has been selected to cover all trip requests that have been considered so far (i.e., the union of fixed requests,  $\bar{R}_t$ , that will not be considered for optimization and those postponable ones,  $\bar{R}_t$ , that would possibly be rerouted and rescheduled in future THs). Generally, as  $t$  increases, the ratio  $\frac{|\bar{R}_t|}{|\bar{R}_t|}$  increases, and a lower percentage of requests are changeable. TH  $t$  also has a set of new requests that have never been considered, denoted as  $R_t$ . Therefore, at the heart of the optimization problem for TH  $t$  is to generate additional plans  $\Omega_t$  and select a subset of plans from  $\bar{\Omega}_t \cup \Omega_t$  to ensure each request  $r \in \bar{R}_t \cup \bar{R}_t \cup R_t$  is covered exactly once. The mathematical formulation for this problem is presented in the appendix.

**Inner Decomposition via Column Generation.** The previous set partitioning problem can be solved very efficiently even when there are millions of plans (or columns) because of its very sparse technical coefficient matrix (Sun et al. 2020). Unfortunately, the set of new plans  $\Omega_t$  is exponentially large and cannot be generated easily because of the combinatorial nature of the vehicle routing problem. A plan  $s \in \Omega_t$  could be generated

by integrating  $r \in \bar{R}_t \cup R_t$  into plans in  $\bar{\Omega}_t$ ; alternatively, a new plan could be created from scratch by exclusively covering riders in  $\bar{R}_t \cup R_t$ . Column generation is employed to strategically generate only high-quality plans within a manageable time, thus avoiding exhaustively enumerating all possible plans (Menezes et al. 2010). This is achieved by decomposing the set partitioning problem into a master problem and a subproblem, detailed in the appendix.

After no new plans can be found for TH  $t$ , an integer program is solved by considering all generated plans and the set of optimum plans becomes an input to TH  $t + 1$  until the end of the planning horizon.

### Reoptimization Procedure

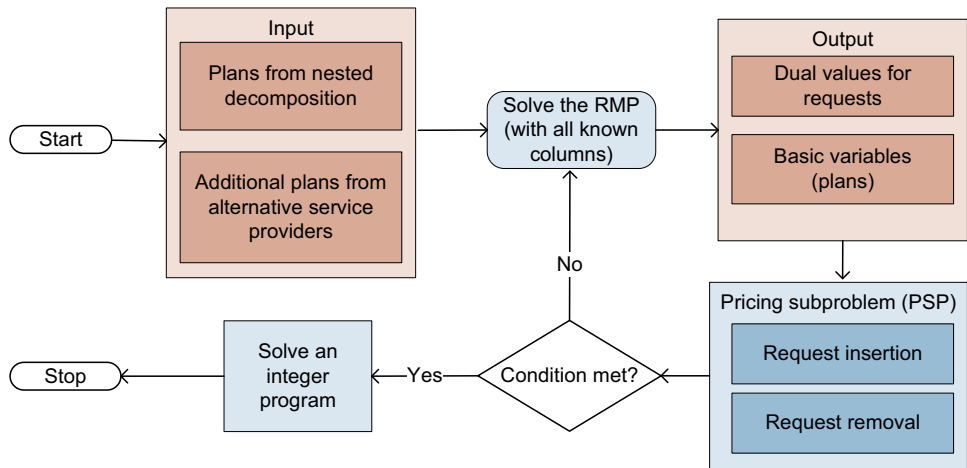
In a dynamic problem where riders are notified and vehicles are dispatched constantly, the conclusion of a rolling horizon procedure means that all final decisions have been made and cannot be reversed. Conversely, the rolling horizon approach in this study is employed to achieve the temporal decomposition of a static paratransit scheduling problem; all decisions made throughout the process can thus be reoptimized, as no commitments have been made to any riders. We explicitly consider the option of serving a trip with alternative service types during the reoptimization procedure.

In light of the column generation-based approach detailed in the appendix, the reoptimization procedure is illustrated in Figure 6 and described as follows. First, all selected plans at the end of the nested decomposition approach are taken as the initial plans. For each individual request, one additional plan is generated, representing the potential option of being served by taxi or Uber. Second, the restricted master problem, a linear program, is solved with all known columns to obtain the dual value associated with each request. In addition, columns in the basis (called basic columns) are also known. Third, two groups of columns are generated in the pricing subproblem (PSP): the first group involves individually inserting requests with positive dual values into each column in the basis if the basic column does not contain the request; then, the second group is generated by removing requests with positive dual values from each basic column if possible. Fourth, after generating two groups of columns, only those with negative reduced costs will be included in the restricted master problem (RMP) to potentially improve the objective. Then, we go back to the second step where the updated RMP is solved, and this process continues until getting no new columns. The final solution is obtained by solving a set partitioning problem, an integer program, with all known columns.

### Results from Real-World Deployments

The new paratransit optimization method described previously has been integrated into the MRMS and

**Figure 6.** (Color online) The Reoptimization Procedure Iteratively Solves an RMP and a PSP to Improve the Plans from Nested Decomposition

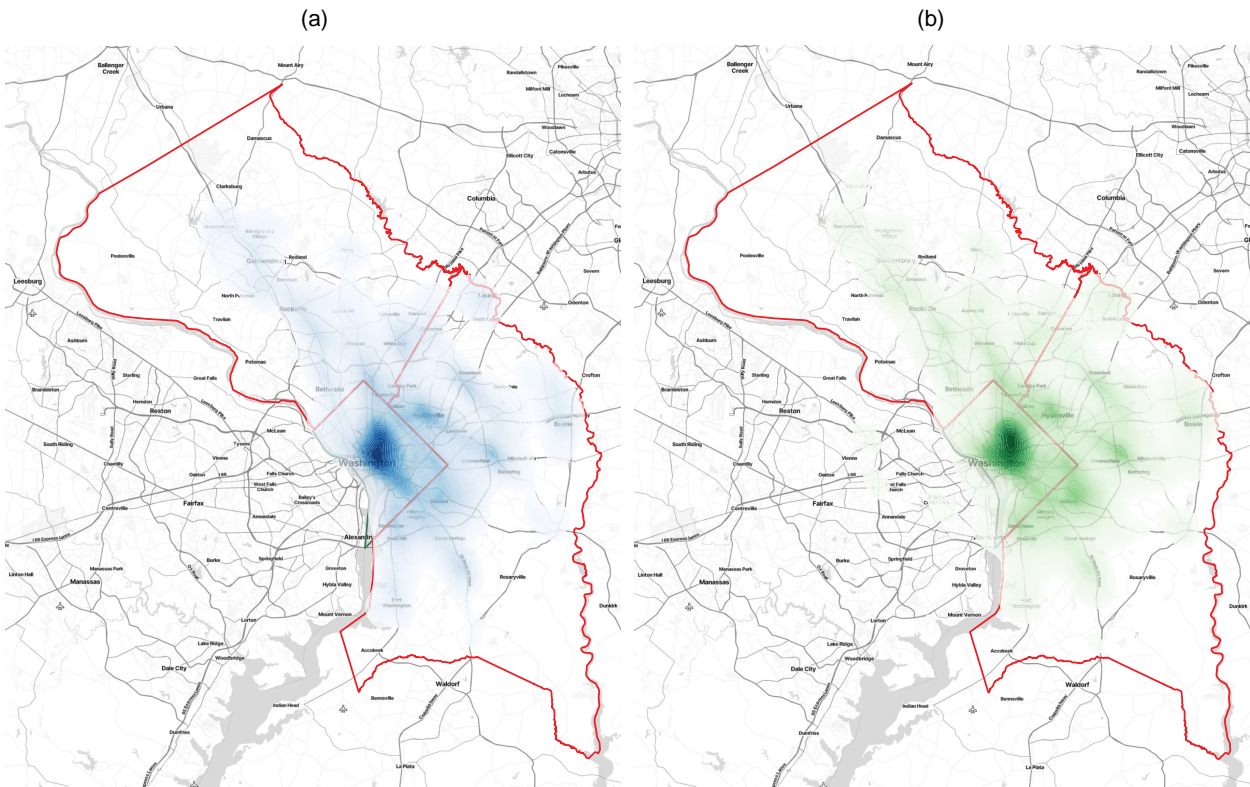


made available to Challenger Transportation for real-world implementation in its area of service. In this section, we compare the legacy algorithm with the new optimization method, using the same constraints and parameter settings, such as scheduling time window, vehicle capacity, and maximum ride time, for a fair comparison. A period of five weekdays (January 8–12)

in 2024 is selected. The daily number of trip requests varies from 1,420 to 1,574. Figure 7 further visualizes the density of trip origins and destinations over the five-day period.

We first examine various operational metrics achieved each day by both optimization approaches. Thanks to the nested decomposition approach and reoptimization

**Figure 7.** (Color online) The Spatial Distributions of Pickup and Drop-off Locations Indicate the High Trip Density in Downtown Washington, DC



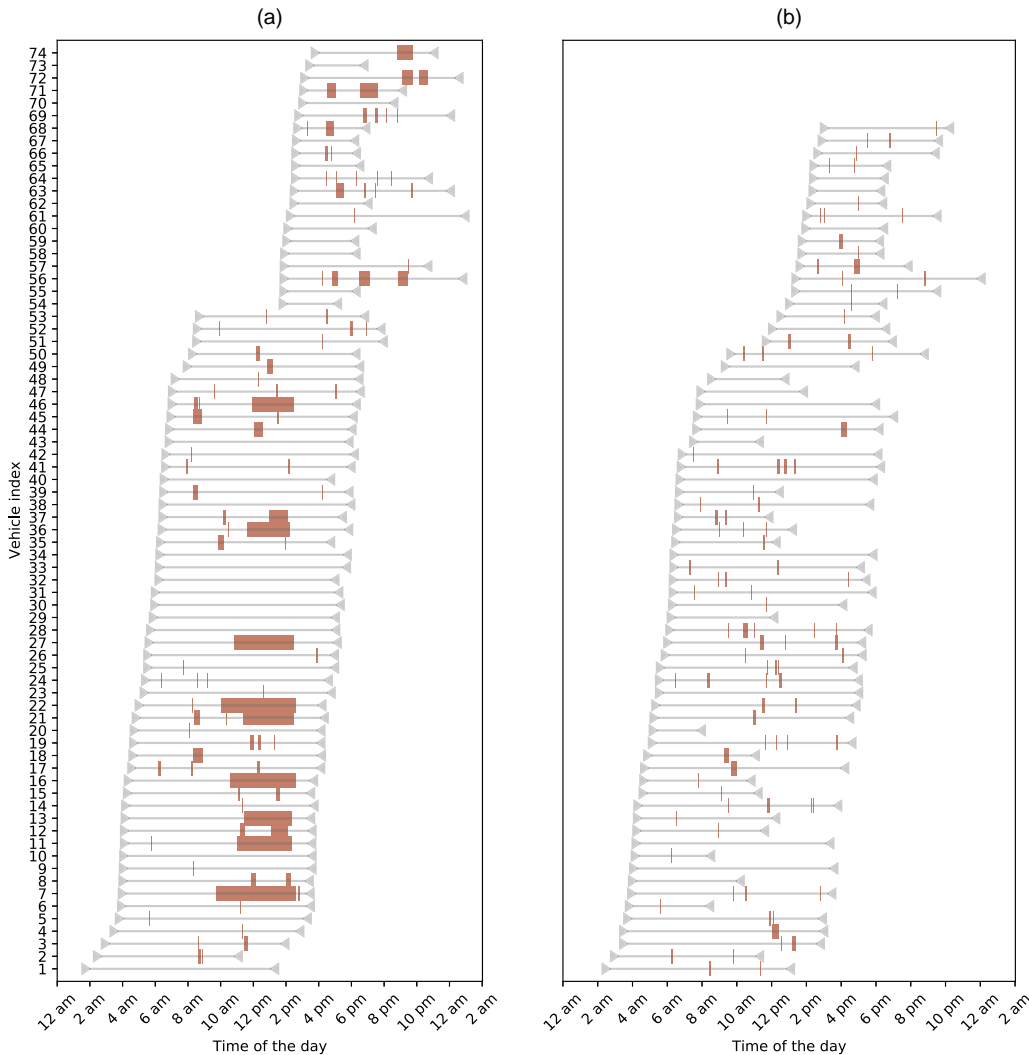
Notes. (a) Pickup location density. (b) Drop-off location density.

procedure, the current optimization approach outperforms the legacy approach in every metric by a large margin. Specifically, the average reductions in total driving distance, driving time, and vehicle idle time are 22%, 24%, and 89%, respectively. Figure 8 shows the vehicle schedules for both optimization methods (legacy versus current), using the January 12 demand data as an example. A vehicle could leave the depot as early as 2 a.m.; the depot return time could be past midnight. The average route duration is 9.3 hours for the legacy approach and 7.8 hours for the current approach, although fewer routes are generated using the current approach. More importantly, Figure 8 highlights a major difference in vehicle idling periods: a substantial reduction in idle time achieved by the current approach. This is attributed to the inherent shortcoming of the PIH, which often results in excessive idling in vehicle schedules. When

inserting a new trip, MRMS’s original PIH prioritizes insertions into existing routes and creates a new route only when no feasible insertions are found for any existing routes.

It is worth noting that the current approach requires 67% less computation time than the PIH because of the effectiveness of nested decomposition. On average, 7% of trip requests, which are “outlier” trips, are determined to be assigned to alternative mobility options, which is essential to ensure overall operational efficiency. Figure 9 shows the temporal patterns of trips, namely, trips in predawn, during rush hours, and at night, are more likely to be assigned to alternative modes. Those riders transferred to the ARP experience a travel time reduction of around 10 minutes per trip. Finally, the number of vehicle routes needed to fulfill all the trip requests, shown in the last column of Table 1, is

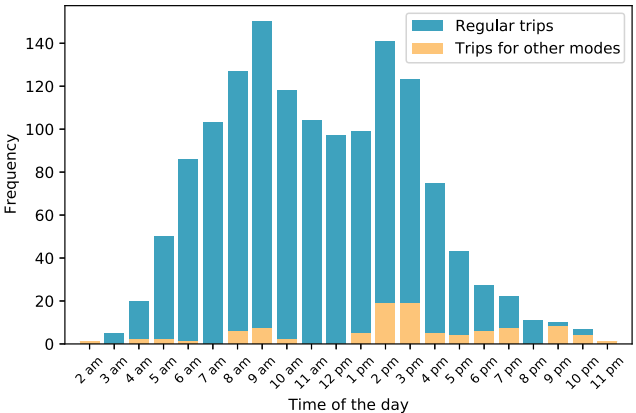
**Figure 8.** (Color online) The Route Start and End Times for Each Vehicle, Along with Vehicle Idling in Between, Under Both Approaches



Notes. (a) By legacy approach. (b) By current approach. Start and end times are represented by gray triangles, and vehicle idling is marked as red bars.



**Figure 9.** (Color online) The Temporal Distribution of Both Groups of Trips Reveals That Trips Served by Alternative Options (i.e., Through the ARP) Are Usually in the Early Morning, Peak Hours, and Near Midnight



comparable across the two approaches. This is because the optimization objective is to minimize the total operating cost rather than the number of scheduled vehicle routes.

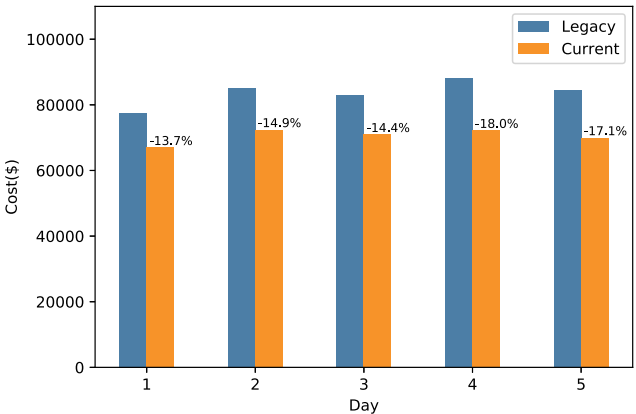
We then analyze the monetary savings. Figure 10 shows the daily total service delivery cost before and after implementing the next-generation paratransit optimization method. It shows that the total operating cost has decreased by, on average, \$13,078 daily, or 15.6%. The annual operating cost savings for CT are estimated at \$4.5 million.

If this optimization approach is adopted by other paratransit contractors of MetroAccess in the entire DC metropolitan area, the annual cost reductions are estimated to be \$19 million to \$34 million, based on the CT’s possible market share of 14.0%–25.0%.

Lessons Learned

In this collaboration, many ideas and thoughts have been exchanged between the university and industry teams during weekly discussions in the last three years. In addition to the successful modernization and deployment of the MRMS, some useful insights and lessons have emerged, which are elaborated next.

**Figure 10.** (Color online) The Daily Operating Cost of Challenger Transportation Was Reduced by ~15.6% After Applying the Current Paratransit Optimization Method



Bridging the Gap Between Academia and Industry

The evolution of paratransit scheduling and routing algorithms represents a journey from the early days of basic heuristics to the current era of computationally intensive methods. Initially, the parallel insertion heuristic was recognized as a pioneering approach to optimizing paratransit routes, developed during a time when computational resources were relatively scarce (Toth and Vigo 1996). The heuristic continues to maintain a strong presence in some commercial software suites today. Nevertheless, the substantial advancements in computing power witnessed over recent decades have facilitated the adoption of more sophisticated algorithms. The shift prompts the need for a critical reevaluation of operational frameworks to fully harness contemporary computational advancements to enhance performance.

Nonetheless, the translation of the latest algorithmic advancements into practical applications is not without obstacles, considering the divide between academic pursuits and industry needs. Academic research often prioritizes theoretical elegance and algorithmic innovation, although it may not adequately address the pragmatic requirements of real-world applications.

**Table 1.** Comparison of Operational Metrics in the Pilot Study of Five Days

Approach	Day	Driving distance (miles)	Driving time (hours)	Idle time (hours)	ARP trips ( <i>n</i> )	Routes ( <i>n</i> )
Legacy	1	17,294	683	48	0	74
	2	19,018	751	53	0	77
	3	18,024	739	67	0	76
	4	19,051	780	83	0	78
	5	18,284	742	79	0	77
Current	1	13,466 (−22.1%)	529 (−22.5%)	7 (−85.1%)	99	68
	2	14,662 (−22.9%)	577 (−23.1%)	7 (−86.9%)	107	73
	3	14,635 (−18.8%)	574 (−22.3%)	6 (−91.7%)	101	76
	4	15,116 (−20.7%)	590 (−24.3%)	8 (−90.8%)	75	77
	5	13,687 (−25.1%)	536 (−27.8%)	7 (−91.3%)	124	69

For example, although advancements in exact algorithms and new valid inequalities offer performance improvements such as a reduced optimality gap, their practical utilization is hindered by significant computational resource needs and a lack of scalability. To illustrate this, the algorithm proposed by Liu et al. (2015) takes four hours to solve instances with only 22 trip requests. Moreover, the accessibility of advanced algorithms to industry practitioners presents a critical challenge. Theoretical constructs, including complex optimization methodologies and mathematical theories, often exceed the comprehension of many industry professionals who are experienced in modern software development but lack familiarity with optimization algorithms. The success of new routing algorithms in enhancing paratransit services hinges on their comprehensibility and applicability to industry contexts. Additionally, exact algorithms typically require licenses for commercial solvers such as CPLEX or Gurobi, but industry practitioners may hesitate to incur additional costs associated with these licenses. The financial implications of purchasing and maintaining licenses for advanced computational tools can be significant, particularly for organizations operating under tight budget constraints. Overall, bridging the gaps between theoretical innovation and practical application, particularly in terms of computational efficiency and ease of usage, is paramount for realizing the potential of advanced routing algorithms in transforming paratransit services.

### Leveraging Recurring Trip Patterns for Efficient Scheduling

A paratransit operator typically serves a finite number of registered riders whose travel needs follow predictable patterns. For example, a rider has to travel to a dialysis facility according to a fixed schedule, repeating weekly or biweekly. Such riders' frequent destinations and appointment times are typically recorded in the operator's database. Paratransit practitioners have long recognized the potential benefits of identifying and seek to leverage these recurring trip patterns to enhance scheduling efficiency. By recognizing these repetitive trip requests, operators can optimize routes by preserving certain segments that have been repeatedly proven as efficient.

We believe that integrating heuristic insights derived from these recurring trip patterns can substantially improve the efficiency of the scheduling process. Rather than constructing routes from scratch each day, a labor-intensive and time-consuming task, the new approach involves modifying previously successful routing plans or consolidating similar trips into distinct route segments. To the authors' best knowledge, such a promising approach has not been explored in the scholarly literature.

### Opening Paratransit Doors to the General Public

A key factor contributing to the subpar operational efficiency of paratransit is its low demand density, which limits the degree of ridesharing. Making underutilized paratransit vehicles available to the general public can improve trip density, thus leading to efficiency gains. For instance, when non-ADA riders complete their trips at a metro rail station, their last legs can be served by ADA paratransit vehicles, as advocated by Plano et al. (2020). A successful model demonstrated by the National Express in Massachusetts blends fully accessible paratransit vehicles with microtransit services, which caters to the population with disabilities and simultaneously meets the general transportation needs of the wider public, irrespective of ADA qualifications (Mey 2022). As a result, more passengers are served per hour. Clearly, opening unused paratransit capacities to the general public can effectively fill service voids, thereby enhancing the overall efficiency and reach of public transportation. This paradigm change, however, needs support from transportation planners and policymakers.

### Rethinking Fleet Capacity and Utilization

Another issue in paratransit lies in the disparity between anticipated and actual vehicle occupancy rates. By design, vehicles have a physical capacity ranging from nine to 11 passengers, as reflected in most paratransit optimization studies in the literature. Nonetheless, many vehicles typically carry only two to three passengers on average in practice. Consequently, a considerable number of seats remain consistently unoccupied. This mismatch has prompted certain operators to transition from using vans to sedans, with the aim of enhancing operational efficiency and cost-effectiveness (Rodman et al. 2024). For example, WMATA has opted to replace vans with sedans or ramp-equipped minivans, a strategic move aimed at addressing maintenance-related issues and ensuring sustained fleet reliability (WMATA 2023b). This adaptation not only reflects a pragmatic response to observed service patterns but also indicates the necessity of ongoing reassessment of operational assumptions in light of real-world performance metrics.

### Conclusion

The academia-industry partnership supported by the NSF to enhance the WMATA's MetroAccess service represents a significant advancement in addressing the operational and financial challenges faced by ADA paratransit. This study introduced a nested decomposition approach combined with a reoptimization procedure to optimize large-scale paratransit scheduling and routing problems. The new approach has been deployed in the MRMS of IT Curves and implemented by Challenger Transportation in its operations. The

deployment resulted in substantial improvements in operational efficiency, including a reduction in operating costs by at least 15%. Some lessons learned during the collaboration were also summarized in this paper.

The successful collaboration reaffirms the capabilities of our advanced optimization algorithms to enhance the financial sustainability and service quality of paratransit. The promising research outcomes not only benefit the Washington, DC, metropolitan area of implementation but also set a precedent for similar improvements in other metro areas, thus promoting a more efficient, cost-effective, and user-centric paratransit service landscape across the United States.

Future research could focus on exploiting the wealth of data. In the realm of paratransit, digital transformation has unlocked the potential for leveraging vast amounts of data through predictive analytics to enhance operational efficiency and service quality. Specifically, using the data, the operator could analyze patterns, predict future demand, and preemptively address potential challenges, such as by actively allocating resources and planning routes for potential demand surges (Qin et al. 2020, Azagirre et al. 2024). However, challenges such as the integration of data analytics into organizational practices, investment in data infrastructure, and promotion of academia-industry collaborations for the development and application of advanced analytical tools will be key areas to address.

Further improving solution quality is a promising direction when computational resources are less constrained. The current temporal decomposition approach is built upon the premise that trips during similar periods can be merged into efficient route segments; however, spatially similar trip requests could also be organized into equally efficient route segments. Therefore, supplementing the current temporal decomposition approach with a spatial decomposition method could yield routes of higher quality, although an increase in total computation time is expected.

Lastly, it is a promising direction to further improve the solution quality when computational resources are less constrained. The current temporal decomposition approach is built upon the belief that trips in similar periods could be merged into efficient route segments; however, spatially similar trip requests could be organized into equally efficient route segments. Therefore, supplementing the current temporal decomposition approach with a spatial decomposition method would yield routes of higher quality, although an increase in the total computation time is expected.

## Acknowledgments

The financial support by the National Science Foundation is gratefully acknowledged, but it implies no endorsement of the findings.

## Appendix

### Mathematical Formulation for the Set Partitioning Problem

With notation introduced in the section titled Set Partitioning Problem, the mathematical program in time horizon  $t$  can be formulated as

$$\text{Minimize } \sum_{s \in \Omega_t \cup \Omega_t} \rho_s z_s \quad (\text{A.1a})$$

$$\text{s.t. } \sum_{s \in \Omega_t \cup \Omega_t} \varphi_{sr} z_s = 1 \quad \forall r \in \dot{R}_t \cup \vec{R}_t \cup R_t \quad (\text{A.1b})$$

$$z_s \in \{0, 1\} \quad \forall s \in \dot{\Omega}_t \cup \Omega_t. \quad (\text{A.1c})$$

The Objective (A.1a) minimizes the total operating cost, where  $\rho_s$  represents the operating cost of route  $s$ . The decision variable  $z_s$  indicates whether plan  $s \in \dot{\Omega}_t \cup \Omega_t$  is selected or not in the solution. The parameter  $\varphi_{sr}$  is one if a request  $r$  is covered in plan  $s$ , and zero otherwise. Constraints (A.1b) ensure that every request  $r \in R_t$  is satisfied exactly once. Constraints (A.1c) restrict the domain of decision variables  $z_s$ .

### Column Generation Approach

It is impractical to enumerate all plans in  $\Omega_t$  for any realistic problem instances. More importantly, it is unnecessary to identify them all because only a tiny portion of the plans in  $\Omega_t$  are eventually selected in the optimal solution. Hence, a promising approach is to iteratively generate high-quality plans using column generation.

The set partitioning formulation is decomposed into an RMP and a pricing subproblem (PSP). The RMP is “restricted” because only a subset of plans in the set partitioning problem is considered. Additionally, Constraint (A.1c) is relaxed as

$$0 \leq z_s \leq 1, \quad \forall s \in \dot{\Omega}_t \cup \Omega_t, \quad (\text{A.2})$$

which results in a linear program that can be solved efficiently.

The main idea of the column generation approach is to first solve the RMP based on a set of initial columns. Then, the dual value for each constraint becomes available and is passed to the PSP. Based on the dual value information, a PSP aims to find better plans with a negative reduced cost, which indicates the potential of improving the objective value of RMP. The RMP is then updated with these plans and solved again to improve the objective. The process described previously is repeated until no profitable plans can be found or another criterion is reached. The final solution is obtained by solving an integer version of RMP based on all obtained solutions.

A step-by-step description is provided as follows:

- Step 1: Find existing plans for fixed requests. Plans in  $\dot{\Omega}_t$  covering requests in  $\dot{R}_t \cup \vec{R}_t$  are the initial columns in the RMP for TH  $t$ . For TH  $t$ , plans  $\dot{\Omega}_t$  are those selected at the end of the previous TH to cover all requests considered before TH  $t$ . For the initial TH, the set of existing columns could be empty.

- Step 2: Generate new columns for changeable requests. We generate a subset of feasible plans covering requests in  $\vec{R}_t \cup R_t$ . One recommended way is to enumerate those plans covering a limited number of requests, as the enumeration



can be done in minimal time because of the few ways to serve the involved riders by one vehicle. For instance, there is only one way to serve a rider, which is to simply pick up and then drop off the rider. When two riders are involved, there are six possible pickup and drop-off sequences. As the number of riders to be served per vehicle route increases, the number of possible pickup and drop-off sequences grows exponentially, which implies that a complete enumeration approach is no longer viable.

Following a similar study by Sun et al. (2020), we limit the number of requests per vehicle in the initial columns to two. In other words, we generate one-request and two-request plans as initial solutions.

A one-request plan involves only one request to be served by the vehicle. Specifically, such a route  $(0+, r+, r-, 0-)$  is generated for any request  $r \in \bar{R}_t \cup R_t$ , where  $0+$  and  $0-$ , respectively, represent the departure and return depots. By design, this one-request plan satisfies all routing constraints, such as maximum ride time, maximum route duration, and pickup time window, because a vehicle is dispatched to serve a rider exclusively. Although such a route is almost impossible to be eventually selected for implementation, it serves as a foundation for other promising routes.

Then, two-request plans, which involve two different requests, are generated. For each pair of requests  $r$  and  $r'$ , where  $r \neq r'$  and  $r, r' \in \bar{R}_t \cup R_t$ , six possible routes are generated, namely,  $(0+, r+, r-, r'+, r'-, 0-)$ ,  $(0+, r+, r'+, r-, r'-, 0-)$ ,  $(0+, r+, r'+, r'-, r-, 0-)$ ,  $(0+, r'+, r+, r-, r'-, 0-)$ ,  $(0+, r'+, r+, r'-, r-, 0-)$ , and  $(0+, r'+, r'-, r+, r-, 0-)$ .

The feasibility of generated plans is checked using the three-pass procedure in Hunsaker and Savelsbergh (2002) in which the vehicle schedule would be determined as a byproduct. Specifically, the first pass checks the route's capacity and time window constraints. The arrival time and departure time of each node are adjusted to the earliest and latest possible ones, respectively; the second pass adjusts the timings from the route's end to its start to shift pickup times later by reducing any available waiting time; and the third pass rechecks the route from start to end to confirm that the time adjustments made previously do not result in any new violations of constraints, such as the maximum ride times and time windows.

The generated one-request and two-request plans are added to the RMP.

- Step 3: Solve the RMP to obtain dual values. The RMP considering all newly added columns is solved by linear programming solvers such as CPLEX to obtain the dual value  $\pi_r$  of each request  $r$ .

- Step 4: Search for promising columns by solving the PSP. A new column is promising in improving the RMP objective if the reduced cost is negative, namely,

$$\rho_s - \sum_{r \in \bar{R}_t \cup \bar{R}_t \cup R_t} \varphi_{sr} \pi_s < 0. \quad (\text{A.3})$$

To achieve the best possible results, we can insert all requests in  $\bar{R}_t \cup R_t$  into all known plans and only keep the promising plans. However, considering the significant computation

time needed, this is not practical in the case of large-scale problems. Instead, an insertion heuristic can be adopted from Sun et al. (2020) for generating those columns. If a plan  $s$  is in the basis, its reduced cost must be zero and is more likely to become negative after subtracting a nonnegative dual value (namely, inserting another request). Therefore, only the plans with  $z_s > 0$  (i.e., in the basis) are to be inserted. Second, only requests with positive dual values ( $\pi_s > 0$ ) are considered for insertion. This is because other requests (with  $\pi_s = 0$ ) cannot help decrease the reduced cost, as can be seen in Equation (A.3). Hence, all requests with  $\pi_s > 0$  are inserted into the plans, with  $z_s > 0$  in the cheapest locations and with the optimal schedules. Additionally, in this process, only those requests in  $\bar{R}_t \cup R_t$  can be inserted into other plans because only those requests can be optimized regarding their routes and schedules at TH  $t$ . The precedence orders of the planned requests  $\bar{R}_t$  in planned routes  $\bar{\Omega}_t$  are unchanged. A more detailed description can be found in Sun et al. (2020). Then, return to Step 3 until no promising columns can be generated.

- Step 5: Obtain the final solution. As a final step, an Integer Program (A.1a)–(A.1c) is solved by including all columns generated.

**Table A.1.** List of Symbols and Abbreviations

Symbol	Meaning
$t$	The index of a time horizon
$r$	The index of a rider request
$R_t$	The set of new requests in time horizon $t$
$\bar{R}_t$	The set of fixed requests in time horizon $t$
$\bar{R}_t$	The set of postponable requests in time horizon $t$
$s$	A vehicle routing and scheduling plan (or simply plan)
$\bar{\Omega}_t$	The set of plans that cover all trip requests that have been considered so far in time horizon $t$
$\Omega_t$	The set of additional plans to further cover new requests in time horizon $t$
$\rho_s$	The operating cost of route $s$
$\varphi_{sr}$	A binary parameter that takes the value of one if a request $r$ is covered in plan $s$ ; zero otherwise
$\pi_s$	The dual value of trip $s$
$z_s$	Binary variable; $z_s = 1$ if plan $s$ is selected; zero otherwise
ADA	Americans with Disabilities Act
APTA	American Public Transportation Association
ARP	Abilities-Ride Program
CT	Challenger Transportation
DARP	Dial-A-Ride Problem
MRMS	Mobile Resource Management System
NSF	National Science Foundation
RMP	Restricted master problem
RP	Rolling period
PDP	Pickup and delivery problem
PIH	Parallel insertion heuristic
PSP	Pricing subproblem
SIH	Sequential insertion heuristic
TCRP	Transit Cooperative Research Program
TH	Time horizon
WMATA	Washington Metropolitan Area Transit Authority

## References

- Azagirre X, Balwally A, Candeli G, Chamandy N, Han B, King A, Lee H, et al. (2024) A better match for drivers and riders: Reinforcement learning at Lyft. *INFORMS J. Appl. Analytics* 54(1): 71–83.
- Bruck BP, Castegini F, Cordeau J-F, Iori M, Poncemi T, Vezzali D (2020) A decision support system for attended home services. *INFORMS J. Appl. Analytics* 50(2):137–152.
- CBO (2024) Federal support of public transportation operating expenses. Accessed March 1, 2024, <https://crsreports.congress.gov/product/pdf/R/R47900>.
- Chen S, Rahman MH, Marković N, Siddiqui MIY, Mohebbi M, Sun Y (2024) Schedule negotiation with ADA paratransit riders under value of time uncertainty. *Transportation Res. Part B Methodological* 184:102962.
- Chicago Metropolitan Agency for Planning (2023) Funding paratransit in Northeastern Illinois. Accessed March 1, 2024, [https://cmap.illinois.gov/wp-content/uploads/PART\\_recommendations-c2-paratransit.pdf](https://cmap.illinois.gov/wp-content/uploads/PART_recommendations-c2-paratransit.pdf).
- Hanne T, Melo T, Nickel S (2009) Bringing robustness to patient flow management through optimized patient transports in hospitals. *Interfaces* 39(3):241–255.
- Ho SC, Szeto WY, Kuo Y-H, Leung JM, Petering M, Tou TW (2018) A survey of dial-a-ride problems: Literature review and recent developments. *Transportation Res. Part B Methodological* 111: 395–421.
- Hunsaker B, Savelsbergh M (2002) Efficient feasibility testing for dial-a-ride problems. *Oper. Res. Lett.* 30(3):169–173.
- Jaw J-J, Odoni AR, Psarftis HN, Wilson NH (1986) A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. *Transportation Res. Part B Methodological* 20(3): 243–257.
- Kessler DS (2004) Computer-aided scheduling and dispatch in demand-responsive transit services. Transit Cooperative Research Program (TCRP) Synthesis 57, Transportation Research Board, Washington, DC.
- Liu M, Luo Z, Lim A (2015) A branch-and-cut algorithm for a realistic dial-a-ride problem. *Transportation Res. Part B Methodological* 81:267–288.
- Luo Z, Liu M, Lim A (2019) A two-phase branch-and-price-and-cut for a dial-a-ride problem in patient transportation. *Transportation Sci.* 53(1):113–130.
- Marković N, Nair R, Schonfeld P, Miller-Hooks E, Mohebbi M (2015) Optimizing dial-a-ride services in Maryland: Benefits of computerized routing and scheduling. *Transportation Res. Part C Emerging Tech.* 55:156–165.
- Menezes F, Porto O, Reis ML, Moreno L, de Aragão MP, Uchoa E, Abeledo H, do Nascimento NC (2010) Optimizing helicopter transport of oil rig crews at Petrobras. *Interfaces* 40(5):408–416.
- Mey N (2022) What is paratransit and why is it important? Accessed March 1, 2024, <https://sparelabs.com/en/blog/what-is-paratransit>.
- Mohebbi M, Miller-Hooks E, Schonfeld P (2023) Management system for dial-a-ride operations. Accessed March 1, 2024, <https://itcurves.net/management-system-for-dial-a-ride-operations/>.
- Planners Collaborative, Inc. (2007) WMATA final report. Accessed March 1, 2024, [https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/WMATA\\_Para\\_Final\\_Report\\_070626.doc](https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/WMATA_Para_Final_Report_070626.doc).
- Plano C, Behrens R, Zuidgeest M (2020) Toward evening paratransit services to complement scheduled public transport in Cape Town: A driver attitudinal survey of alternative policy interventions. *Transportation Res. Part A Policy Practice* 132:273–289.
- Qin Z, Tang X, Jiao Y, Zhang F, Xu Z, Zhu H, Ye J (2020) Ride-hailing order dispatching at DiDi via reinforcement learning. *INFORMS J. Appl. Analytics* 50(5):272–286.
- Rahman MH, Chen S, Sun Y, Siddiqui MIY, Mohebbi M, Marković N (2023) Integrating dial-a-ride with transportation network companies for cost efficiency: A Maryland case study. *Transportation Res. Part E Logist. Transportation Rev.* 175:103140.
- Rist Y, Forbes MA (2021) A new formulation for the dial-a-ride problem. *Transportation Sci.* 55(5):1113–1135.
- Rodman W, Etminani-Ghasrodashti R, Blume K, Edrington S, Tung L-W (2024) *Paratransit Fleet Configurations* (The National Academies Press, Washington, DC).
- Shrode G (2022) The mass transit fiscal cliff: Estimating the size and scope of the problem. *Eno Transportation Weekly* (September 23), <https://enotrans.org/article/the-mass-transit-fiscal-cliff-estimating-the-size-and-scope-of-the-problem/>.
- Sun Y, Chen Z-L, Zhang L (2020) Nonprofit peer-to-peer ridesharing optimization. *Transportation Res. Part E Logist. Transportation Rev.* 142:102053.
- Transit Cooperative Research Program (2018) ADA paratransit service models. Accessed March 1, 2024, <https://nap.nationalacademies.org/catalog/25092/ada-paratransit-service-models>.
- Toth P, Vigo D (1996) Fast local search algorithms for the handicapped persons transportation problem. Osman IH, Kelly JP, eds. *Meta-Heuristics* (Springer, Boston), 677–690.
- Uber (2023) Case study: WMATA meets its goals by giving paratransit riders more choices. Uber (March 27), <https://www.uber.com/blog/case-study-wmata-giving-paratransit-riders-more-choices/>.
- WMATA (2021) FY2022 proposed budget effective July 1, 2021. Accessed March 1, 2024, <https://www.wmata.com/about/records/upload/Proposed-FY2022-Budget.pdf>.
- WMATA (2022a) Application for MetroAccess door-to-door paratransit service for people with disabilities. Accessed March 1, 2024, [https://www.wmata.com/service/accessibility/metro-access/upload/MetroAccessApplicationFillable\\_accessible\\_HQ\\_05-24.pdf](https://www.wmata.com/service/accessibility/metro-access/upload/MetroAccessApplicationFillable_accessible_HQ_05-24.pdf).
- WMATA (2022b) FY2023 budget effective July 1, 2022. Accessed March 1, 2024, <https://www.wmata.com/about/records/upload/FY2023-Approved-Budget-Final.pdf>.
- WMATA (2023a) Customer guide to MetroAccess. Accessed March 1, 2024, <https://www.wmata.com/service/accessibility/metro-access/upload/MetroAccess-Customer-Guide.pdf>.
- WMATA (2023b) Performance report FY 2023. Accessed March 1, 2024, [https://www.wmata.com/about/records/upload/MetroPerformanceReport\\_FY23Q4\\_1Report\\_20230922.pdf](https://www.wmata.com/about/records/upload/MetroPerformanceReport_FY23Q4_1Report_20230922.pdf).
- Zhen L, Lv W, Tan Z, Dong B (2022) Container transportation scheduling between port yards and the hinterland in Yunfeng. *INFORMS J. Appl. Analytics* 52(3):250–266.

## Verification Letter

Pierre Matabaro, Vice President of Operations, Challenger Transportation, 8210 Beechcraft Avenue, Gaithersburg, Maryland 20879, writes:

“It is my pleasure to verify the successful implementation of the optimization method and software documented in the paper titled ‘Empowering MetroAccess Service with Nested Decomposition and Service Type Integration’ by Chen et al.

“Challenger Transportation (CT) is a MetroAccess operator that has been providing paratransit services in the District of Columbia, Maryland, and Virginia since 2000. As we operate nearly 80 vehicles and serve approximately 1,500 riders daily, we employ the Mobile Resource Management System (MRMS), a computerized paratransit scheduling software suite developed by IT Curves, for vehicle scheduling and routing. The routing quality directly affects our service delivery cost and profit. As an MRMS user for many years, we have been pleased with the vehicle routes built by MRMS;

nonetheless, we do occasionally notice certain deficiencies that could be addressed, such as extensive vehicle idling in some routes.

“IT Curves started a collaboration with Florida State University and University of Utah researchers on upgrading the MRMS in late 2021. CT was invited to their weekly meetings to share our practical experience as an end user. The university team conducted a systematic diagnosis of the existing scheduling algorithm and developed a brand-new algorithm, which was proved to generate superior performance. Through numerous rounds of testing and validation, IT Curves incorporated the new algorithm in MRMS, which was then made available to us. CT has been using the upgraded MRMS since November 2023. We find that the MRMS 2.0 can cut the total operating cost by 10%–18%. We are especially pleased with how MRMS can intelligently filter out trips for taxis, which makes a huge difference in consolidating routes for efficiency. Those results reported in this paper were from the second week of January 2024. We estimate that CT can save above \$4 million yearly.

“We anticipate continued collaborations with IT Curves and the university researchers, whose innovative work can lead to higher operational efficiency and customer satisfaction. Please do not hesitate to contact me if you require further information.”

**Shijie Chen** is a PhD candidate in the Department of Industrial and Manufacturing Engineering at Florida State University. His main research areas are modeling, analysis, and optimization in transportation systems. Everything about operations research interests him.

**Md Hishamur Rahman** is currently a PhD student and graduate research assistant in the Department of Civil and Environmental Engineering at the University of Utah. His research focuses on optimizing paratransit and public transport services by leveraging data science, machine learning, and operations research.

**Nikola Marković** is an assistant professor of civil and environmental engineering at the University of Utah. His research uses operations research and data science to improve the efficiency of transportation systems. He was a recipient of the 2015 Glover-Klingman Prize and the 2022 INFORMS Transportation Science and Logistics Society Best Paper Award.

**Muhammad Imran Younus Siddiqui** is the chief technology officer and director of research and development at IT Curves. He holds an MS in computer science, having conducted in-depth studies of advanced databases and algorithm analysis and design.

**Matthew Mohebbi** is the founder and CEO of IT Curves. He received his MS in electrical engineering from George Washington University in 1982. In the transit industry, He has focused on industry engineering and product development, assisting in the programming of IT Curves’s patented routing algorithms.

**Yanshuo Sun** is an associate professor of industrial engineering at the FAMU-FSU College of Engineering. He specializes in transportation systems optimization. He received his PhD in civil engineering from the University of Maryland, College Park.