Joint design of fixed-route and paratransit services with autonomous pods

Xiaoyu Yan[a], Hongyu Zheng[b], Tianxing Dai[a], Yu (Marco) Nie[a,*]

[a] *Department of Civil and Environmental Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA*

[b] *Department of Industrial and Systems Engineering, University of Tennessee, Knoxville, TN 37996, USA*

Abstract

This study envisions a jointly designed transit system comprised of a fixed-route (FR) service and a paratransit (PT) service. The integration of the two services is inspired by the potential application of modular autonomous vehicles, or pods, in transit. Constrained by a fixed budget, the operator of the joint system aims to minimize the total user cost by optimally allocating pods between the two services. To formulate the operator's design problem, we propose a stylized model, in which the FR service features a simple 2D grid route structure overlaying on a square city, and the PT service is designed as a general on-demand system that can be configured in different modes of operations. A case study is conducted using transit data from the Chicago region. We find that joint design helps prevent resource misallocation that could render a service dysfunctional under insufficient budgets, although its potential to reduce total user cost is limited. Enforcing the equal-access constraint—requiring that PT users incur no greater cost than FR users—tends to help PT users at the expense of FR users, though the overall impact on total user cost is insignificant. Modularity enables the formation of pod trains using small pods, which benefits FR operations, particularly when the design is not tightly constrained by budget. In contrast, automation delivers greater service improvements for PT users, whose more labor-intensive cost structure makes them more sensitive to efficiency gains, especially under tight budgets. Among the PT service modes, ridesharing is the most flexible, allowing for a wide range of service levels based on the available budget.

Keywords: modular autonomous vehicle, fixed-route transit, paratransit, joint design

1. Introduction

Paratransit (PT), mandated by the Americans with Disabilities Act (ADA) of 1990, provides shared-ride, door-to-door and transit feeder services to disabled individuals who have no adequate access to fixed-route services. While the ADA did not create PT, the mandate has significantly stimulated the growth of its use around the country. Because PT services require disproportionately high subsidies, their rapid expansion

*Corresponding author
Email address: y-nie@northwestern.edu (Yu (Marco) Nie)

presents a major challenge to transit agencies. With a fairly small customer base, PT services in 2013 costed $5.2 billion, or roughly 12.2 percent of the entire transit budget; in comparison, its share was 3.2% in 1988, before the ADA was enacted (Kane et al., 2016). In 2012, the average cost of a PT trip was $29.30, which is almost four times of what costs to provide a regular, fixed-route (FR) trip ($8.15) (United States Government Accountability Office, 2012). The cost is even higher in large metropolitan areas. For instance, the average cost of a PT trip was over $70 in New York City in 2016 (Kaufman et al., 2016). In Chicago, it was about $56 in 2024, of which more than 90% must be subsidized by taxpayers[1]. There is clearly a pressing need to make PT services more efficient and financially sustainable.

Previous studies (Kane et al., 2016; Kaufman et al., 2016) suggest that innovative technologies might hold the key to help current PT systems better meet their challenges. One idea is to take advantage of the burgeoning on-demand ride-hail services such as Uber and Lyft. These new services may be integrated with the FR service as a first- and last-mile solution (Jaffe, 2015), or hired as an independent contractor to cover certain service areas or types for which FR services are not cost effective. An example of the latter strategy is the Rideshare Access Program (RAP) and Taxi Access Program (TAP) recently rolled out by the Pace Bus in Chicago, which essentially offer to subsidize an on-demand ride ordered by an eligible PT user[2]. Another development that promises to revolutionize transit operation is automation. Given the exceedingly high share of the labor cost in the operating budget of the large U.S. transit agencies[3], the introduction of driverless vehicle technology in transit could have a dramatic effect on its cost structure, design/operation principles and user experience. A particularly noteworthy invention is modular autonomous vehicles (Lin et al., 2023; Zhang et al., 2020), also known as autonomous pods (referred to as pods hereafter). Pods can run on existing road infrastructure by themselves, individually or collectively as a connected train. Automation improves cost efficiency by reducing the labor cost. Modularity, on the other hand, enhances operational flexibility. According to the U.S. Department of Transportation, the average load factor of the FR bus systems in the U.S. fell below 10% in 2018[4]. Thus, the seat capacities of these systems are severely underutilized, especially during the off-peak periods. The use of pods can mitigate this excessive waste, because they are much smaller (designed to carry just a handful of passengers) than regular buses (with 40 seats or more) and can be linked together to cope with a higher demand when necessary. Moreover, given their modest size, pods may also be used to operate PT services. It is to the study of jointly designing an FR bus service and a PT service with autonomous pods that the present study is dedicated.

Because PT and FR services target different customer bases, they tend to be designed and operated separately in practice. In the Chicago region, for example, the Chicago Transit Authority (CTA) is responsible

[1]See https://www.pacebus.com/budgets, last visited on 7/25/2024

[2]See https://www.pacebus.com/riders-disabilities, last visited on 7/25/2024.

[3]In the Chicago region, the labor cost accounts for nearly 70% of the operating budget, see https://www.transitchicago.com/finance/, last visited on 7/25/2024.

[4]See https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/ntd/data-product/134406/2018-ntst-appendix_0.pdf, last visited on 7/25/2024.

for operating an FR service, while the PT service is delegated to the Pace Bus which is also responsible for the FR service in the region's suburbs. In fact, much of Pace Bus's PT service is operated by specialized contractors such as National Express Transit (NEXT)[5]. Equipped with the modular autonomous vehicle technology, we may finally break the operational barriers that have hitherto separated the two services.

This study envisions a transit operator who acquires a homogeneous fleet comprised of only pods, which are used to support both FR and PT services. The difference is, while a single pod is sufficient as the basic service unit for the PT service, the number of pods needed to form such a unit may increase with the demand for the FR service. Constrained by a budget that determines the size of the pod fleet, the operator aims to minimize a total user cost by first allocating the pods between the two services and then optimally configuring each service. To formulate the operator's design problem, we propose a stylized model built on a square city. While the FR service features a simple 2D grid route structure overlaying on the city, the PT service is designed as a general on-demand system that can be configured in different modes of operation as in Daganzo & Ouyang (2019a). Our objective is to gauge the benefits of the integration enabled by autonomous pods. Based on empirical data from Chicago, our case study will explore the impact of modularity, automation and the mode of paratransit operation on the performance of transit services in a wide range of settings, including both independent and joint designs.

The rest of the paper is organized as follows. The next section briefly reviews the related studies and Section 3 presents the joint design model. Section 4 reports and analyzes the results of a case study constructed using data from the Chicago region. Section 5 concludes the paper with a summary of main findings and recommendation for future research.

2. Related studies

We organize our review of related studies in three subsections: (i) joint transit design aimed at dealing with different types of demands; (ii) transit design with autonomous pods; and (iii) paratransit as on-demand transportation service.

2.1. Joint transit design for heterogeneous demands

Due to the variation of travel demand during the day, there is an inherent imbalance in transit usage during peak and non-peak hours, which, if left unaddressed, could lead to significant waste. A remedy is to augment fixed-route transit services with more flexible operation strategies, such as variable fleet sizes, routes, and time tables. Jara-Díaz et al. (2020) proposed a fixed route design that varies its fleet size between peak and off peak periods. In their design, a primary fleet serves in both peak and non-peak periods, and a

[5]See https://nellc.com/national-express-transit-next-renews-seven-year-paratransit-contract-with-pace/, last visited on 7/25/2024.

secondary fleet operates during the peak period only. They found that this flexibility in fleet management can better accommodate demand fluctuations, leading to improved system efficiency and reliability. Similarly, Liu et al. (2022) found that operating with multiple vehicle types of different capacities helps reduce system cost. However, a potential downside that comes with running heterogeneous fleets is the elevated risks for headway irregularity, which often leads to bus bunching (Jara-Díaz et al., 2020). To address variations in demand across different locations, Ouyang et al. (2014) developed a heterogeneous route configuration that combines a spatially sparser main service with denser services added to neighborhoods of high-demand density. Guo et al. (2021) proposed a flexible bus routing strategy that are responsive to spatiotemporal heterogeneity in travel demand.

2.2. Transit design with autonomous pods

With the introduction of autonomous pods, transit services can better adapt to variable demands. Notably, since pod-based transit systems promise to accommodate demand variations with a homogeneous fleet, they can effectively avoid operational problems like bus bunching (Khan & Menéndez, 2023; Khan et al., 2023; Liu et al., 2024). Researcher have begun to look into the design and operations of such systems. Some of these studies focus on micro-level operating strategies, such as docking, dispatching and transferring (Chen et al., 2021; Chen & Li, 2021; Pei et al., 2021; Wu et al., 2021). Others considered the design at the system level. For example, Liu et al. (2020) attempted to determine the minimum size of a pod fleet, while Liu et al. (2021) aimed to use pods to design a flexible-route transit service. Zhang et al. (2020) designed a door-to-door shuttle service with pods, which consist of individual pods (trailer modules) that can travel locally to serve demand and connect travelers to pod-trains (main modules) for long-distance trips. Khan & Menendez (2024) proposed a fixed-route bus service with pods that by design do not stop. Passengers board or alight a pod-train through a service module that disjoins and rejoins the train at stops. Those who need to transfer do so in-vehicle and in real-time by moving between pods. Cheng et al. (2024) analyzed a similarly seamless, transfer-free bus system under both homogeneous and heterogeneous demand scenarios, though their design is not entirely stop-less. Zheng et al. (2024) also considered a pod-based, fixed-route transit system, though their focus was to analyze co-modality mechanisms through which pods can be shared between a transit operator and a freight carrier.

2.3. Paratransit as on-demand transportation service

Paratransit is a publicly owned on-demand transportation service. Daganzo & Ouyang (2019a) provided a general analysis framework for such services. Gupta et al. (2010) aimed to improving the operational efficiency of paratransit in the Twin Cities of Minneapolis and Saint Paul by testing two strategies. The first is to frequently re-optimize the routes used by the contracted service provider and the other is to selectively use taxi as an alternative. The first strategy was found to be more effective, promising up to

5% savings in the operating cost. Nguyen-Hoang & Yeung (2010) estimated demand and cost functions for paratransit services in the U.S. Their benefit-cost analysis showed that the demand for paratransit is elastic and that the benefits of paratransit "far exceed the costs". Miah et al. (2020) examined the barriers and opportunities for paratransit users to adopt on-demand micro-transit service such as Via. Fu & Ishkhanov (2004) addressed the issue of determining the optimal fleet size and mix (i.e., the share of different type vehicles) for operating a particular paratransit service. The attempts to jointly design paratransit and fixed-route transit service are relatively rare. A recent exception is Rodriguez-Roman et al. (2024), who proposed a design model that aims to minimize a system cost dependent on the level of services of both paratransit and fixed-route services, as well as a measure of inequality in access to the transit services. The total budget must be allocated to the two services so as to minimize the system cost. Using a continuum approximation method first proposed in Rahimi et al. (2014), they measured the level of paratransit service exclusively by spatial coverage around the fixed routes, i.e., the size of the service area, which is modeled as a polygon.

2.4. Research gaps

Our reading of literature indicates that the studies on the joint design of fixed-route and paratransit services are quite limited. Moreover, while the application of autonomous pods in transit has attracted much attention lately, few had considered taking advantage of the flexibility afforded by the technology to integrate paratransit with fixed-route services. This study intends to address these gaps.

3. Model

We consider a square city with a grid road network, as shown in Figure 1. The side length of the grid network is denoted as L. The city's travel demand is divided into two categories: the regular demand and the paratransit demand. The former includes anyone who is not eligible to use paratransit, the latter includes those who are. A transit agency operates a fixed-route (FR) bus service and a paratransit (PT) service in the city. For simplicity, we assume that the origins and destinations of both regular and paratransit demands are uniformly distributed across the city. The trip generation rates per unit area per revenue hour for the regular and paratransit demands are denoted as λ_R and λ_D, respectively. In general, $\lambda_D \ll \lambda_R$. In Chicago, for example, $\lambda_D/\lambda_R \sim 1\%$. Eligible paratransit users are often physically challenged and thus require some support for boarding and alighting. For this reason, paratransit users are assumed to only use the PT service. Because each type of travelers only use the service designated for them in this study, we shall refer to regular and paratransit users as, respectively, FR users and PT users hereinafter.

Both services are operated with a shared fleet of modular autonomous vehicles, or pods. When employed for the FR service, each pod can hold up to b_R passengers, which are limited by the physical size of the pod. For the PT service, the maximum number of passengers allowed in each pod is b_D. We assume $b_D < b_R$

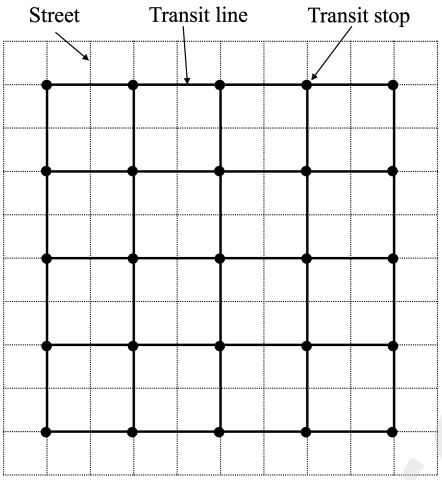


Figure 1: A square city with a grid road network and the fixed-route transit service.

because (i) PT users often have special needs (e.g., wheelchair) that require more space and (ii) too many passengers sharing the same vehicle in an on-demand mode can cause excessive delays.

3.1. Fixed-route (FR) service

The FR service operates in both North-South (N-S) and East-West (E-W) directions. The transit operator determines the number of transit lines in each direction, denoted by N, and the uniform headway, denoted by H. For simplicity, we treat an intersection of a pair of N-S and E-W lines as a stop, which means the stop spacing is given by L/N. An FR user is assumed to always use the stop closest to their origin/destination for access. For those who need transfer to reach their destination, we require that they make at most one transfer to minimize the penalty associated with transfer (e.g., extra waiting time). In a grid network, the expected number of transfers for the FR user can be estimated as $e = \frac{(N-1)^2}{N^2}$ (Chen & Nie, 2017). Note that all trips with a transfer can be executed in two different ways, i.e., first N-S then E-W, or vice versa. In this study, we assume that travelers choose the initial direction randomly. The vehicle used to operate the FR service is a pod-train, or p-train, which consists of several pods coupled together. Although in theory a p-train can be joined or dis-joined en-route for greater efficiency and flexibility (Tian et al., 2022; Chen et al., 2021), the present study does not allow such possibilities. Nor does it allow transfers to take place between pods in the same train as envisioned in Khan & Menendez (2024). Thus, our model might slightly underestimate the benefits of the technology.

One way to formulate the transit operator's design problem is to minimize the user cost while requiring the agency cost not exceed a given budget. We next specify how these costs are related to the decision variables N and H. Following the convention, both are measured for an operating hour (or revenue hour).

The agency cost includes the amortized acquisition cost of pods, and the costs consumed to cover the distance traveled and staff time. For a transit line, the maximum possible number of passengers on a p-train

can be estimated as $\frac{L^2\lambda_R}{8}H\left(\frac{1}{N}+\frac{1}{N-1}\right)$, attained at the center of the line (Cheng et al., 2024). To minimize the agency cost, each p-train should be operated exactly at the capacity when the load factor is the highest. Due to the integer nature of the number of pods, the train size, defined as the number of pods in each p-train, is rounded up to

$$s = \left\lceil \frac{L^2\lambda_R}{8b_R}H\left(\frac{1}{N}+\frac{1}{N-1}\right)\right\rceil. \tag{1}$$

For each round trip, a p-train travels $2L$. There are $2N$ transit lines in total (N lines for each direction). Given the service's headway H, the total vehicle distance traveled by all p-trains within an hour is:

$$Q = \frac{4NL}{H}. \tag{2}$$

The total p-train time per operating hour includes three parts: the time required to traverse the distance at the cruising speed v_0, the total time lost per stop due to deceleration and acceleration, and the total time lost to passenger boarding/alighting. Let v denote the average operating speed. We have

$$\frac{Q}{v} = \frac{Q}{v_0} + \frac{4\tau_1 N^2}{H} + \tau_2(1+e)\lambda_R L^2, \tag{3}$$

where τ_1 is the average lost time per stop and τ_2 is the average boarding/alighting time per passenger. Note that $\frac{Q}{v}$ represents the fleet size, i.e., the number of p-trains required to operate the fixed-route service with N and H. Accordingly, the total number of pods needed is defined as $m_R \equiv sQ/v$. Thus, the hourly agency cost for the FR service, denoted as u_R^a, can be calculated by

$$u_R^a = c_m m_R + c_Q\frac{vm_R}{s} + (c_{dr}+c_M)\frac{m_R}{s}, \tag{4}$$

where c_m is the acquisition cost per pod amortized to each revenue hour, c_{dr} is the hourly driver cost for human-driven vehicles (equals 0 if the pod is operated autonomously), c_Q and c_M are the operating costs associated with unit p-train distance and unit p-train hour, respectively. Here c_M account for the time-based operating costs that cannot be eliminated even if the pod train can drive itself — such costs include but are not limited to computing, software, surveillance, and administrative functions. To capture the aerodynamic benefits from linking the pods, we model c_Q as a concave and increasing function of the train size s, i.e.,

$$c_Q = c_D s^\gamma, \gamma \in (0,1), \tag{5}$$

where c_D is the cost of operating a single pod per unit distance and γ is the parameter related to the aerodynamic benefits. Thus, as more pods are added, a p-train achieves greater efficiency. Specifically,

doubling the size of the train will not increase the distance-based cost by a factor of 2, but by a factor of $2^\gamma < 2$ — the closer γ to zero, the greater the discrepancy.

We further assume all three unit costs increase with the pod size b_R, i.e.,

$$c_m = f_m(b_R), c_D = f_D(b_R), c_M = f_M(b_R), \tag{6}$$

where $f_m(\cdot), f_D(\cdot)$ and $f_M(\cdot)$ are functions to be estimated based on data. In general, we expect these to be monotonically increasing functions.

The total user cost per hour can be estimated as

$$u_R^u = \lambda_R L^2 \left(\frac{H}{2}(1+e) + \frac{L}{N v_w} + \frac{0.34L(2N^2+2N+1)}{N^2 v} \right), \tag{7}$$

where v_w is walking speed of an average passenger. The per capital user cost consists of three parts, corresponding to the three terms in the parenthesis in Equation (7). The first term is related to the average waiting time incurred at the origin stop and the transfer stop; the second is related to the access time (i.e., the average time taken to walk to/from the nearest stop from/to a rider's origin/destination); and the last is related to the in-vehicle time, i.e., the average time spent on riding on a p-train.

Finally, the average cost incurred by an FR user is

$$\bar{T}_R = \frac{u_R^u}{\lambda_R L^2}. \tag{8}$$

3.2. Paratransit (PT) service

The paratransit service is envisioned as an on-demand service that qualified travelers can request in real time. Such a service can be operated in different modes. Taxi is perhaps the oldest and most well-known, providing a direct, door-to-door service for each individual rider. Transportation network companies (TNC) such as Uber and Lyft also provide taxi-like service. A more cost-effective mode is ridesharing, which first emerged in the 1970s in the form of Dial-a-Ride service (Wilson et al., 1976; Stein, 1978). Modern ridesharing implementations include a wide variety of TNC products, each with its own sharing rules.

Similar to the FR service, we assume the operator of the PT service aims to minimize the hourly user cost subject to a budget constraint. The PT operator considers two decisions: the pod fleet size m_D and the operating mode $o \in \mathbb{O} = \{\texttt{TX}, \texttt{DR}, \texttt{RS}\}$, where `TX` refers to taxi, `DR` refers to Dial-a-Ride, and `RS` refers to ridesharing. Note that our definition of the three modes is adopted from Daganzo & Ouyang (2019a); see below for details. To reduce unnecessary detours when multiple users share one vehicle, we require all pods to operate individually for the PT service, and hence there are at most b_D PT users on board.

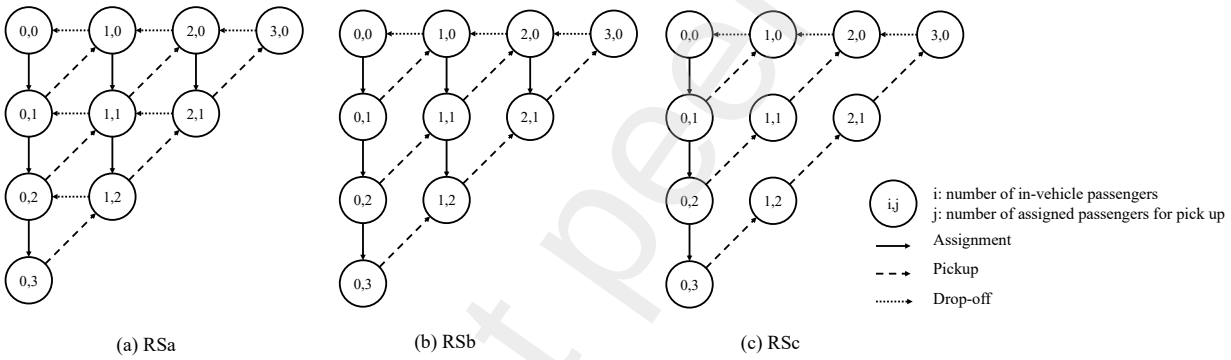
Following Daganzo & Ouyang (2019a) we model the PT service as a workload transition network tied to a mode of operation, which defines such features as pod assignment, pickup, and drop-off. In the workload

transition network, the current condition of the system is represented by a state vector $\boldsymbol{n}$, which consists of the number of pods classified according to workloads (represented by nodes in the network). A pod's workload is described by a tuple of non-negative integers (i, j), where i represents the number of passengers onboard, and j indicates the number of passengers scheduled for pickup. Clearly, the sum of i and j should be less than the pod's capacity b_D. Therefore, we write $\boldsymbol{n} = \{n_{ij}\}$ where n_{ij} is the number of pods at the workload (i, j). Given the pod fleet size m_D, we always have the following relation from the conservation of fleet size

$$m_D = \boldsymbol{n}\mathbf{1} = \sum_{i,j} n_{ij}. \tag{9}$$

Below, we first discuss how to specify such a transition network for ridesharing, before turning to taxi and Dial-a-Ride.

3.2.1. Ridesharing



(a) RSa (b) RSb (c) RSc

Figure 2: Workload transition networks for the rideshairng (`RS`) mode with capacity $b_D = 3$. (a) `RSa`, (b) `RSb` and (c) `RSc` each corresponds to a slightly different operating rule.

Figure 2 shows the workload transition networks for three slightly different rideshare modes. Mode `RSc` in the figure is taken from Daganzo & Ouyang (2019a) whereas Mode `RSa` and Mode `RSb` add more flexibilities. In general, a workload transit network contains $(b_D + 1)(b_D + 2)/2$ nodes. The number of links is $3b_D(b_D + 1)/2$, $b_D(b_D + 2)$, and $b_D(b_D + 5)/2$, respectively, for Mode `RSa`, `RSb` and `RSc` in Figure 2 . As illustrated in the figure, the transition between different workloads can be achieved in three ways, distinguished as solid, dashed, and dotted links. Specifically,

1. an *assignment* event by which the workload (i, j) becomes $(i, j + 1)$, as a new passenger is assigned to the pod for pickup. The assignment rate for workload (i, j) is denoted as a_{ij} [veh/hr];
2. a *pickup* event by which the workload (i, j) changes to $(i + 1, j - 1)$, as the completion of the pickup adds one passenger to the vehicle and removes one from the pickup list. The pickup rate is denoted as p_{ij} [veh/hr].

3. a *drop-off* event, by which the workload (i, j) changes to $(i-1, j)$. The dropoff rate is denoted by d_{ij} [veh/hr].

The links can be represented by three vectors corresponding to the rate of the above events: $\boldsymbol{a}(\boldsymbol{n}) = \{a_{ij}\}, \boldsymbol{p}(\boldsymbol{n}) = \{p_{ij}\}, \boldsymbol{d}(\boldsymbol{n}) = \{d_{ij}\}$. These vectors of event rates are related to the state vector $\boldsymbol{n}$, as well as the topology of the transition network — the latter determined by the operating mode o. Take the three modes illustrated in Figure 2 for example. In Mode `RSa` (the left plot), drop-off and assignment can take place at any workload: a pod may drop off a passenger no matter how many unfilled pickup assignments it has, and a pod may be given new assignment no matter how many passengers are already on-board. For Mode `RSb`, a restriction is imposed to forbid drop-off when there is at least one unfilled pickup assignment. Mode `RSa` is the most restrictive of the three, which, on top of the restriction added in Mode `RSb`, further requires no assignment may be given to a pod that already has passengers on-board.

Denote the link-node incidence matrices of each link type as $\boldsymbol{A}$, $\boldsymbol{P}$, $\boldsymbol{D}$. At the steady state of the system, $\boldsymbol{n}$ must satisfy the following flow conservation condition:

$$\boldsymbol{a}(\boldsymbol{n})\boldsymbol{A} + \boldsymbol{p}(\boldsymbol{n})\boldsymbol{P} + \boldsymbol{d}(\boldsymbol{n})\boldsymbol{D} = \boldsymbol{0}. \tag{10}$$

Thus, given a fleet size m_D, which is a decision variable, the number of pods in each workload level n_{ij} can be solved from the equation system consisting of (9) and (10).

We proceed to show how the operating mode affects the relationship between the transition vectors and the state vector. For Mode `RSa`, a pod can accept new pickup assignments as long as the number of passengers on-board plus passengers already assigned is less than the capacity of the pod. Thus, the total number of pods ready for pickup can be estimated as

$$y = \sum_{i+j<b_D} n_{ij}. \tag{11}$$

Accordingly, in the cases of Figure 2, i.e., when $b_D = 3$, the assignment rate a_{ij}, assume the pickup and drop-off can be completed immediately, the pickup rate p_{ij}, and the drop-off rate d_{ij} are given by

$$a_{ij}(\boldsymbol{n}) = \lambda_D L^2 n_{ij}/y, \quad (i,j) = (0,0),(0,1),(0,2),(1,0),(1,1),(2,0), \tag{12a}$$

$$p_{ij}(\boldsymbol{n}) \approx n_{ij}/(\delta(y) + t_b), \quad (i,j) = (0,1),(0,2),(0,3),(1,1),(1,2),(2,1), \tag{12b}$$

$$d_{ij}(\boldsymbol{n}) \approx n_{ij}/(\delta(i) + t_a), \quad (i,j) = (1,0),(2,0),(3,0),(1,1),(1,2),(2,1). \tag{12c}$$

Equation (12a) states that the total demand should be assigned to a workload (i, j) in proportion to the number of pods at that workload; Equation (12b) states the number of pods at a workload (i, j) can be estimated by the Little's formula, i.e., equal to the product of the pickup flow rate $p_{ij}(\boldsymbol{n})$ and the pickup

time. We posit that the pickup time consists of two parts. The first is the expected travel time required to reach a random location given the pickup fleet size y. This travel time can be estimated by

$$\delta(y) \approx kL/(v_D\sqrt{y}), \tag{13}$$

where k is a constant related to network topology (Daganzo, 1978), which approximately equal to 0.63 in a rectangular network like ours, and v_D is the average operation speed of the pods for the PT service (Daganzo & Ouyang, 2019b). The second part concerns the time between the arrival of an RS vehicle and the boarding of the passenger. For a standard taxi or Uber ride, this time is usually negligible. However, since many paratransit riders are physically challenged, they need extra time to reach and board the vehicle. The term t_b in Equation (12b), which is an exogenous parameter, is introduced to account for this need. Also based on the Little's formula, Equation (12c) states the number of pods at a workload (i, j) is the product of the drop-off flow rate and the average drop-off time — the latter depends on i, the number of passengers on-board at the steady state. Here again, the drop-off time consists of the standard travel time $\delta(i)$ and the average time needed for the passenger to alight the vehicle, denoted as t_a.

With slight modifications, the equation system that mirrors the workload transition can be applied to Modes `RSb` and `RSc` in Figure 2. When applied to `RSb`, workloads (1,1), (1,2), and (2,1) should not appear in Equation (12c). That is, no drop-off flow rates should originate from those work loads. As for Mode `RSc`, in addition to the above change, two more modifications are needed. First, no assignment flows originate from workloads (1,0), (2,0), and (1,1). Thus, these nodes should be removed from Equation (12a). Second, the number of pods ready for pickup assignment should be given as follows instead, since the assignment event is prohibited for any pods with passengers on board:

$$y = \sum_{j<b_D} n_{0j}. \tag{14}$$

When b_D takes a different value (e.g., 2), the above formulation remains valid, though the details will change.

3.2.2. Taxi and Dial-a-Ride

Taxi (Mode `TX`) and Dial-a-Ride (Mode `DR`) can be viewed as special instances of the ridesharing modes. Figure 3 shows the workload transition networks for taxi and Dial-a-Ride modes.

Since no sharing is allowed in the taxi mode, an empty pod (at workload $(0, 0)$) would first get an assignment to transition to $(0, 1)$. Then, after picking up a passenger (arriving at the workload $(1, 0)$), it makes a delivery and transitions back to $(0, 0)$.

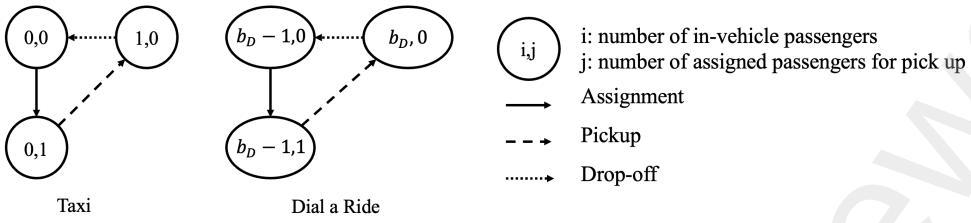


Figure 3: Workload Transition networks for taxi (Mode `TX`) and Dial-a-Ride (Mode `DR`).

In this case, the assignment rate a_{00}, pick-up rate p_{01}, and drop-off rate d_{10} are given by

$$a_{00}(\mathbf{n}) = \lambda_D L^2, \tag{15a}$$

$$p_{01}(\mathbf{n}) \approx \frac{n_{01}}{\delta(n_{00}) + t_b} \approx \frac{n_{01}}{\frac{kL}{v_D\sqrt{n_{00}}} + t_b}, \tag{15b}$$

$$d_{10}(\mathbf{n}) \approx \frac{n_{10}}{\delta(1) + t_a} \approx \frac{n_{10}}{\frac{kL}{v_D} + t_a}. \tag{15c}$$

The Dial-a-ride mode illustrated in Figure 3 may be viewed as an extreme form of ridesharing that prioritizes productivity over user experience. Under this mode of operation, the goal is to fully utilize the capacity of each pod at the steady state, thereby minimizing the fleet size. To achieve this goal, the pod should always alternate between three work loads: ($b_D - 1$,0), ($b_D - 1$, 1) and (b_D, 0). The iteration starts at ($b_D - 1$,0), at which the pod has $b_D - 1$ passengers on board, and is given a new passenger to pick up. This will bring the pod's work load to (b_D, 0) (fully loaded) before it drops off the next passenger. Daganzo & Ouyang (2019a) noted that this mode tends to reduce the number of pods at work load ($b_D - 1$,0) to zero while keeping some passengers waiting for assignment. Denoting the number of waiting passengers as z, we could estimate the pickup time as $\delta(z)$ and the transition flows are given by

$$a_{(b_D-1)0}(\mathbf{n}) = \lambda_D L^2, \tag{16a}$$

$$p_{(b_D-1)1}(\mathbf{n}) \approx \frac{n_{(b_D-1)1}}{\delta(z) + t_b} \approx \frac{n_{(b_D-1)1}}{\frac{kL}{v_D\sqrt{z}} + t_b}, \tag{16b}$$

$$d_{b_D0}(\mathbf{n}) \approx \frac{n_{b_D0}}{\delta(b_D) + t_a} \approx \frac{n_{b_D0}}{\frac{kL}{v_D\sqrt{b_D}} + t_a}. \tag{16c}$$

3.2.3. Agency and user costs

The agency cost for the PT service also includes three parts: the acquisition cost, the operating cost related to the total distance travelled and that related to hours. It is given by

$$u_D^a = c_m m_D + c_D v_D \left(m_D - \sum_{i+j \le b_D} (t_b p_{ij} + t_a d_{ij}) \right) + (c_{dr} + c_M) m_D, \tag{17}$$

where c_m, c_D, c_{dr} and c_M are the same parameters defined for the FR service. These costs are assumed to be identical to their counterparts in the FR service if the same type of pods are employed. Note that, for the distance cost (the second term), we must deduct the the portion of the fleet tied up in the boarding and alighting process.

The average cost for a PT user can be estimated from the Little's formula as follows:

$$\bar{T}_D = \frac{\sum_{i+j\leq b_D}(i+j)n_{ij}}{\lambda_D L^2}, \tag{18}$$

where the numerator is the number of passengers in the system and the denominator is the hourly demand generation rate. Thus, the total user cost is

$$u_D^u = \lambda_D L^2 \bar{T}_D = \sum_{i+j\leq b_D}(i+j)n_{ij}. \tag{19}$$

In the case of `DR`, since a certain number of users (z) are waiting at any time, their waiting time should be counted, leading to

$$u_D^u = \sum_{i+j\leq b_D}(i+j)n_{ij} + z. \tag{20}$$

3.3. Optimal design problems

We first discuss how each of the two services can be designed separately. For the PT service, the specification of the design model depends on the operating mode o, which is formulated as follows:

$$\min_{m_D,\boldsymbol{n}} \quad u_D^u(m_D,\boldsymbol{n}) \tag{21a}$$

$$\text{s.t.} \quad u_D^a(m_D,\boldsymbol{n}) \leq B_D, \tag{21b}$$

$$m_D \geq \underline{M^o}, \tag{21c}$$

$$\text{Equations (9) and (10) as dicated by mode } o \in \mathbb{O}. \tag{21d}$$

The design objective is to minimize the user cost with a maximum agency budget B_D, by choosing a fleet size m_D and a vector of workloads $\boldsymbol{n}$. Constraint (21c) requires the fleet size be no less than a minimum value, denoted by $\underline{M^o}$ for mode o, needed to maintain a stable service (stable in the sense that the waiting time would not grow indefinitely). For a given demand and mode, $\underline{M^o}$ can be estimated from the workload transition equations (Daganzo & Ouyang, 2019a). Constraint (21d) dictates that the decision variables must also meet the workload transition relationship per the mode. Finally, we note that `RS` $\in \mathbb{O}$ contains the collection of all three sub-modes illustrated in Figure 2.

For the FR service, we formulate the design problem as:

$$\min_{N,H,s} \quad u_R^u(N,H,s) \tag{22a}$$

$$\text{s.t.} \quad u_R^a(N,H,s) \le B_R, \tag{22b}$$

$$H \in [\underline{H}, \overline{H}], N \in [\underline{N}, \overline{N}], s \in [1, \overline{s}]. \tag{22c}$$

Similarly, the objective is to minimize the user cost, subject to a maximum agency budget B_R, by choosing the number of lines N, headway H, and train size s. All three decision variables are restricted by their respective natural lower and upper bounds: headway cannot be too small (which can lead to instability) or too large (for reasonable level of service), the number of lines cannot be too small (or the access distance would be unreasonably long) or too large (constrained by the typical block size), and the p-train cannot be too long (for maneuver feasibility) or too short (each train must have at least one pod). In addition, s is in fact an intermediate variable, whose value depends on N and H according to Equation (1), which also ensures its integrality.

3.4. Joint design of FR and PT services

We are now ready to present a design model that aims to jointly optimize both FR and PT services. The operator is assumed to face the following problem: given a fixed budget, a fleet of autonomous pods is acquired to operate both an FR and a PT service such that (i) the total cost of both regular and paratransit users is minimized and (ii) all operating constraints are observed. For a given PT mode $o \in \mathbb{O}$, the problem can be formulated as follows:

$$\min_{N,H,s,m_D,\boldsymbol{n}} \quad u_R^u(N,H,s) + u_D^u(m_D,\boldsymbol{n}) \tag{23a}$$

$$\text{s.t.} \quad u_R^a(N,H,s) + u_D^a(m_D,\boldsymbol{n}) \le B, \tag{23b}$$

$$H \in [\underline{H}, \overline{H}], N \in [\underline{N}, \overline{N}], s \in [1, \overline{s}], \tag{23c}$$

$$m_D \ge \underline{M^o}, \tag{23d}$$

$$\text{Equations (9) and (10) as dictated by mode } o \in \mathbb{O}. \tag{23e}$$

Here, B is the total hourly budget available to operate both services. Most constraints in the above formulations are transplanted from the individual design models presented in previous sections, and thus require no further explanation.

Another constraint that may be added to the joint design model (23) concerns the equal-access clause in the ADA, which states that passengers who are unable to use the FR service are entitled to equal service within the catchment area of the FR service. In this study, we interpret the clause as requiring the average

travel time experienced by a PT rider should be no greater than that of an FR rider, i.e.,

$$u_R^u(N, H, s)/\lambda_R \geq u_D^u(m_D, \boldsymbol{n})/\lambda_D. \tag{24}$$

Finally, we note that the joint design model(23) still depends on the operating mode o. To find the best mode for a given demand and budget level, the operator may simply solve Problem (23) multiple times, one for each of the five PT modes (recall `RS` has three sub-modes), and chooses the best solution.

4. Case Study

Our experiments contain three parts. In Section 4.2, we report the results of independently designed FR and PT services, focusing on the impact of modularity, automation, and in the case of PT, the mode of operation. Section 4.3 then explore the benefits of joint design and the effect of the equal-access clause. Finally, we perform a sensitivity analysis on a few parameters in Section 4.4. In all experiments, the design model is solved using Python's SciPy package[6]. In the following, we first describe how the models are set up with empirical data.

4.1. Setup

Our case study is constructed using the transit data provided by the Chicago Transit Authority (CTA) and Pace Bus. Table 1 reports the default values of the input parameters for the design models. These values may be obtained (i) from literature (in which case the source is noted), (ii) by the discretion of the modeler (marked as "Selected"), (iii) by estimation based on publicly available data (marked as "Estimated"), and (iv) by a calibration process. We first discuss Category (iii) below.

In 2019, CTA covered a service area of 803 km^2 and served about 237.3 million bus rides (Chicago Transit Authority, 2020a), with a fleet of 1,864 buses and a bus operating cost of $824.3 million (Chicago Transit Authority, 2020b). Thus, we configure the square city in our model to match the CTA service area, which gives $L = 28.3$ km. In the same year, the Pace Bus served about 4.28 million PT rides (Regional Transportation Authority Mapping and Statistics, 2024), of which about 3.24 million rides originated from within the CTA service area. It operates in total 552 vehicles for the PT service, with an annual operating budget of $186.3 million in 2019 (Pace Suburban Bus Service, 2020a). Based on these data, we estimate that, for FR and PT services respectively, the demand rates are 68.8 pax/hr/km^2 and 0.691 pax/hr/km^2, the hourly operating budgets are $\$191,957$ and $\$26,681$, and the numbers of vehicles employed for our case study are 1,864 and 461 (See Appendix A.1 for more details).

[6] `https://scipy.org/`

The boarding time t_b is estimated to be 10 minutes, twice the waiting time permitted under the Comfort Uber service. The alighting time t_a is estimated at t_a=5 minutes.

We next specify the pod operating costs. First, the driver cost c_{dr} is estimated at \$40/hr, based on a base wage of about \$25/hr plus benefits. To obtain those operating costs that vary with seat capacity (see Equation (6)), we assume, without loss of generality, that the seat capacity $b_R \in [6, 50]$ and denote the acquisition costs for $b_R = 6$ and $b_R = 50$, $\underline{c}_m$ and $\overline{c}_m$, respectively, the cost of operating a single pod per unit distance for $b_R = 6$ and $b_R = 50$ as $\underline{c}_D$ and $\overline{c}_D$, respectively, and the non-driver cost of operating a single pod per unit time for $b_R = 6$ and $b_R = 50$ as $\underline{c}_M$ and $\overline{c}_M$, respectively. Assuming the unit cost is a linear function of b_R, we have

$$f_m(b_R) = \underline{c}_m + \alpha_m(b_R - 6), f_D(b_R) = \underline{c}_D + \alpha_D(b_R - 6), f_M(b_R) = \underline{c}_M + \alpha_M(b_R - 6), \forall b_R \in [6, 50], \quad (25)$$

where $\alpha_m = (\overline{c}_m - \underline{c}_m)/44, \alpha_D = (\overline{c}_D - \underline{c}_D)/44$, and $\alpha_M = (\overline{c}_M - \underline{c}_M)/44$.

Table 1: Main parameters and their default values for the design models.

Notation	Meaning	Value	Unit	Source
General				
R	Service area	803	km^2	Estimated
L	Side length	28.33	km	Estimated
c_{dr}	Driver cost per unit vehicle hour	40	$\$/hr/pod$	Estimated
$\overline{c}_m$	Capital acquisition cost of a 50-seat bus amortized to each revenue hour	9	$\$/hr/pod$	Calibrated
$\underline{c}_m$	Capital acquisition cost of a 6-seat pod amortized to each revenue hour	1.5	$\$/hr/pod$	Calibrated
$\overline{c}_M$	Operating cost per unit vehicle hour for a 50-seat vehicle	38	$\$/hr/p_{train}$	Calibrated
$\underline{c}_M$	Operating cost per unit vehicle hour for each p-train with 6-seat pod	9	$\$/hr/p_{train}$	Calibrated
$\overline{c}_D$	Operation cost per 50-seat pod distance	0.8	$\$/hr/km$	Calibrated
$\underline{c}_D$	Operation cost per 6-seat pod distance	0.4	$\$/hr/km$	Calibrated
FR Service				
λ_R	Demand rate for fixed-route (FR) service	68.8	$pax/hr/km^2$	Estimated
B_R	FR peak hour operating budget	191,957	$\$/hr$	Estimated
m'_R	Fleet size of FR service at status quo	1864	veh	Estimated
b_R	Seat capacity of the pod for FR service	[6,12,25]	pax/pod	Selected
$\overline{b}_R$	Seat capacity of the bus for FR route service	50	pax/pod	Selected
H_0	Average CTA bus headway at peak hour	12.5	min	Calibrated
L/N_0	Average CTA stop spacing	0.4	km	Dai et al. (2024)
v_w	Walking speed	2	km/hr	Daganzo & Ouyang (2019a)
v_0	Vehicle cruising speed for the FR service	25	km/hr	Daganzo & Ouyang (2019a)
τ_1	Average lost time due to deceleration/acceleration at each stop	12	$sec/stop$	Chen & Nie (2017)
τ_2	Average boarding/alighting time per regular passenger	1	sec/pax	Chen & Nie (2017)
γ	Aerodynamic energy saving parameter	0.5	-	Selected
$[\underline{H},\overline{H}]$	Lower and upper bounds on headway	[3, 40]	min	Selected
$[\underline{N},\overline{N}]$	Lower and upper bounds on the number of lines	[20, 120]	-	Selected
PT Service				
λ_D	Demand rate for paratransit service	0.691	$pax/hr/km^2$	Estimated
B_D	PT hourly operating budget	26,681	\$/hr	Estimated
m'_D	Fleet size of PT service in Chicago area at status quo	461	veh	Estimated
b_D	Seat capacity for PT riders	2 or 3	pax/pod	Selected
t_b	Boarding time for a PT rider	10	min/pax	Estimated
t_a	Alighting time for a PT rider	5	min/pax	Estimated
v_D	Vehicle cruising speed for the PT service	25	km/hr	$= v_0$
k	Rectangular grid network topology coefficient	0.63	–	Daganzo (2010)

Appendix A.2 details how the parameters in Category (iv)—namely, $\underline{c}_m$, $\overline{c}_m$, $\underline{c}_D$, $\overline{c}_D$, $\underline{c}_M$, and $\overline{c}_M$—are

obtained through a calibration process. We assume that, under the status quo, the FR service operates with 50-seat buses, while the PT service runs with six-seat vans in the DR mode with $b_D = 3$, both employing human drivers. Table 2 compares key operational metrics—including budget, fleet size, ridership, and the labor cost share—derived from the calibrated models with those observed under the status quo. For both services, the budgets in the calibrated models differ by no more than 0.7% from the status quo values. For the FR service, the labor cost share in the budget is approximately 75.5% in the calibrated model, compared to about 77% in the published data (Federal Transit Administration, 2022). Although the labor cost share for the contracted PT service is unknown, given the smaller vehicle size, we expect a higher labor share, which is consistent with the calibration result about (85%).

Table 2: Comparison of status quo and calibration results for FR and PT services.

	Agency Cost ($/hr)	**Ridership** (pax/hr)	**Fleet Size**	**Share of Labor Cost** (%)	**Avg Travel Time** (hr)
FR status quo	191957	55256	1864	77	-
FR Calibration	192921	55256	1867	75.5	1.366
PT Status quo	26681	554.9	461	-	-
PT Calibration (DR3)	26509	554.9	461	85.2	2.525

4.2. Results of independent designs

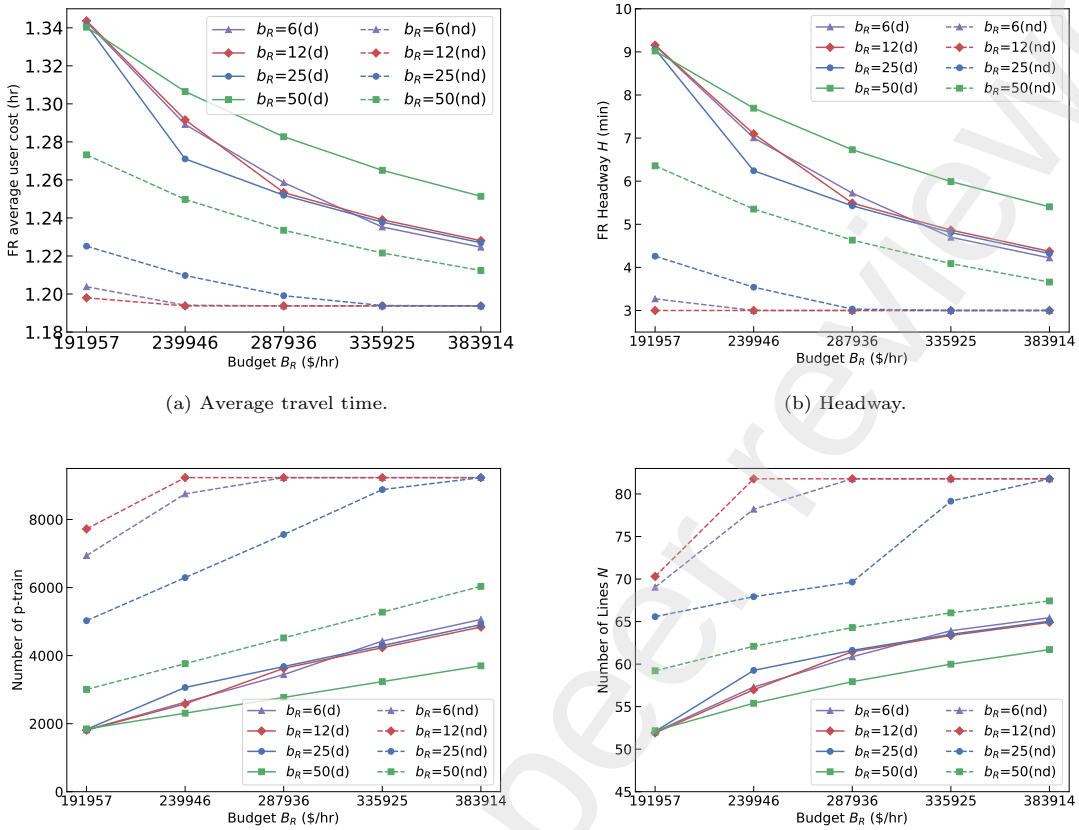
4.2.1. FR service

We examine how key performance metrics vary with budget available, the seat capacity (b_R) and automation. For budget, we start with the baseline level ($191,957/hr) and gradually increase to twice of that value ($383,914/hr) — in total five budget levels are tested. Four seat capacities are considered: 6 (the default value for PT service), 12, 25, and 50 (the default value for FR service). Finally, with automation, we set $c_{dr} = 0$ whereas with drivers $c_{dr} = \$40$/hr by default. Figure 4 reports the results of the experiments.

A general and expected trend is that more budget leads to better service: shorter headway (smaller H), denser coverage (larger N), more p-trains, and lower average travel time. Also expected is that the returns on the investment gradually diminish as the budget increases.

The impact of modularity is complicated by its interaction with automation. With human drivers, using smaller pods offers no advantage at the baseline budget. With high labor costs and a limited budget, it makes sense to operate larger vehicles, which can meet the same demand with longer headways. As the budget increases, smaller pods become more attractive. At 1.25 times and 1.5 times the baseline budget, the optimal pod size is 25, and at higher budget levels, the smallest pod size (6) is optimal. Notably, at the three highest budget levels, there is little difference between using pods with 6, 12, or 25 seats, whereas a pod with 50 seats results in significant efficiency losses.

Automation significantly alters the outcomes. First, it results in higher levels of service. At the lowest budget level, automation reduces headways by a factor of 1.5 to 3 (depending on seat capacity) and increases



(a) Average travel time.

(b) Headway.

(c) Number of p-trains.

(d) Number of lines.

Figure 4: Impact of modularity and automation on the performance of FR service. In the plots, b_R represents seat capacity and *d* and *nd* refer to human-driven (with drivers) and autonomous (no drivers).

coverage density by 20–40%. However, the improvement in average travel time is more modest, at about 5–10%. As the budget increases, these gains become more moderate—partly because abundance offsets the cost-reduction advantage of automation, and partly because a lower bound on headways is enforced at higher budgets to avoid operational instability. Second, automation changes the calculus of modularity. Without drivers in p-trains, as we increase the available budget the design quickly reaches the minimum headway and maximum coverage (where service level is at maximum and budget is in excess) for pod size 12. This implies that pod size 12 utilizes the resources most efficiently. On the other hand, a pod with 50 seats consistently performs the worst. Again, once the budget level exceeds the baseline level, the differences among pods with 6, 12, or 25 seats diminish.

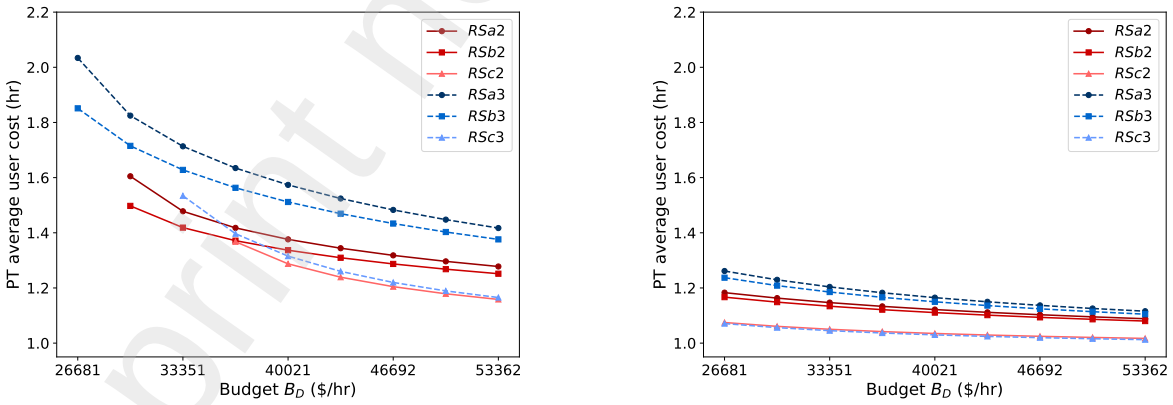
4.2.2. PT service

For PT service, we similarly test multiple budget levels between the baseline ($26,681/hr) and twice the baseline. Note that only six-seat pods are used for PT service.

We first conduct an experiment comparing the performance of the three ridesharing modes presented in Figure 2, referred to as `RSa`, `RSb` and `RSc`. For each mode, we allow either 2 or 3 riders in the pod at the same time. This leads to six ridesharing modes: {`RSa2`,`RSa3`,`RSb2`,`RSb3`,`RSc2`,`RSc3`}. As shown in Figure 5a, higher budgets improve service levels (measured by average user cost), while allowing more riders per pod tends to worsen it. Since `RSa` and `RSb` offer greater operational flexibility—as illustrated by additional links in the workflow diagrams—they require a lower minimum budget to operate. The minimum budget for `RSc` is about $35,000/hr, whereas both `RSa` and `RSb` become operational at budgets below $30,000/hr. However, as the budget increases, `RSc` outperforms the other two modes in terms of the level of service. Thus, there is a clear trade-off between operational efficiency and level of service. The greater the flexibility, the more efficient the operation, the lower the level of service. Another interesting finding is that the number of riders allowed in the pod has a much great impact on the performance of `RSa` and `RSb` than that of `RSc`, likely due to the flexibility in pickup and drop-off operations.

Comparing Figure 5a to Figure 5b confirms that the general trends remain similar with or without automation. However, `RSc` outperforms the other modes across all budget levels. Moreover, automation has a significant impact, especially at lower budget levels. At the baseline, `RSc` achieves an average user cost of about 1.1 hours, compared to nearly 2 hours for `RSb3` at the same budget—an almost 50% reduction. Thus, automation appears to benefit PT service more than FR service (where the largest improvement in average user cost ranges from 5% to 10%). One reason is that, in our setting, driver costs represent a higher share of the labor cost for PT service than for FR service.

Unless otherwise specified, hereafter we use `RSn` (where n is the number of riders allowed) to represent the mode that delivers the best performance among {`RSan`,,`RSbn`,`RScn`}, for $n = 2, 3$.



(a) With drivers (human-driven).

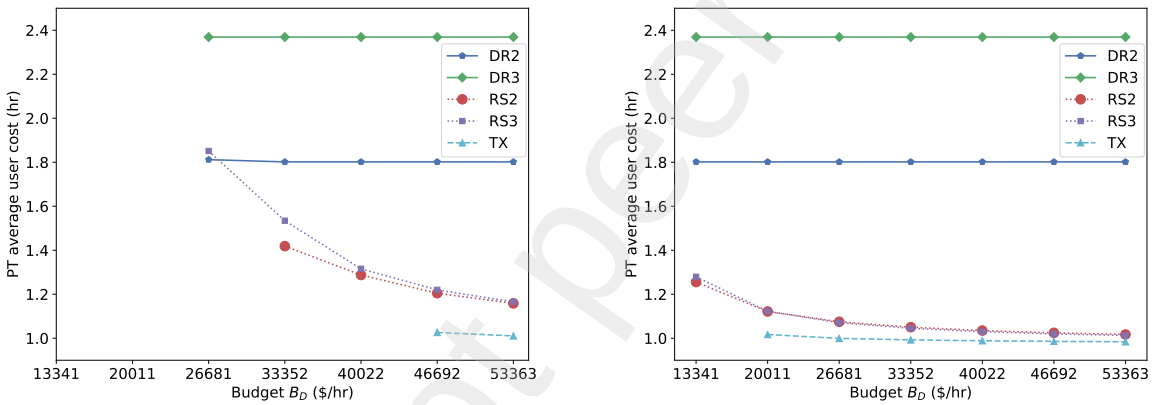
(b) Without drivers (Automation).

Figure 5: Performance of three `RS` modes at different budget levels. For `RSxn`, x refers to different mode (a, b or c), and n refers to the number of riders allowed in the vehicle (2 or 3).

Figure 6 further compares the two best ride sharing modes, `RS2` and `RS3`, against taxi (`TX`) and dial-a-ride

(DR2 and DR3). Here, we let the budget range from 0.5 times the baseline to 2 times the baseline to highlight the differences between these modes.

When human drivers are required (Figure 6a), none of the five modes are operational below the baseline budget level. The two Dial-a-Ride modes, along with RS3, begin functioning at the baseline, whereas TX only becomes operational once the budget exceeds $45,000—nearly double the baseline. Clearly, because taxi offers an exclusive door-to-door service, it demands a higher budget. Dial-a-Ride modes are notable for their insensitivity to additional budget once the minimum requirement is met. This is because, in a Dial-a-Ride mode, vehicles always operate at full delivery capacity, minimizing the fleet size at the expense of service level. In other words, the mode is designed to function at a bare minimum service level, and thus cannot effectively leverage extra funding. In this regard, the two ridesharing modes differ significantly. For example, RS3 performs similarly to DR2 at the baseline budget but outperforms it as the budget increases.



(a) With drivers.

(b) Without drivers.

Figure 6: Performance of five PT modes at different budget levels. For RSn and For DRn, n refers to the number of riders allowed in the vehicle (2 or 3).

As seen from Figure 6b, all modes benefit from automation in that they become operational at lower budgets—even taxi operates at only 0.75 times the baseline budget. Automation does not enhance the level of service for Dial-a-Ride modes, as their service level is inherently limited by the peculiar mode of operation. It does, however, positively impact the two ridesharing modes and, to a lesser extent, taxi. Interestingly, once taxi meets the minimum fleet requirement, there is also very little room for further improvement. The reason is that, beyond a certain threshold, increasing the number of circulating vehicles does not significantly reduce pickup and drop-off times.

4.3. Results of joint designs

In this section, we conduct three sets of design experiments with the two budget levels (baseline and 1.5 times baseline) to gauge the benefits of joint design. The first includes three designs, each integrating an FR

service operated by 50-seat pods with a different PT mode (TX, RS2, RS3) operated by six-seat pods (DR2 and DR3 are inferior so they are excluded from the discussion hereafter). All pods are driven by humans. The second set is similar to the first except that the FR service is operated by six-seat pods, which enables us to explore the benefits of modularity. We create the third set by removing the driver cost from the second set so that we can examine the effect of automation on joint design. For easy reference, we shall label Sets 1, 2 and 3 as J-50-H, J-6-H and J-6-A, respectively. The independent design corresponding to the status quo (50-seat bus for FR service and six-seat van for PT service, all human-driven), labeled as NJ, is used as the benchmark.

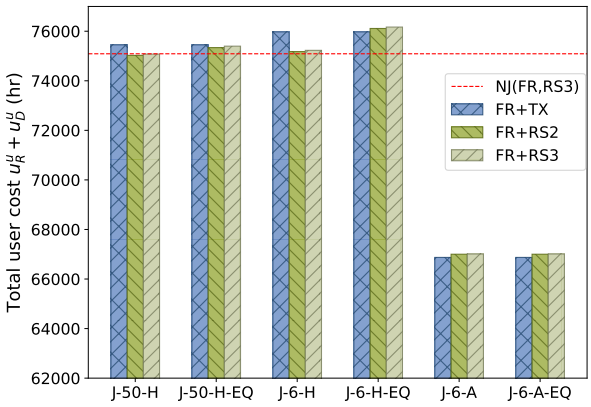
In what follows, Section 4.3.1 reports the results of these experiments without enforcing the equal-access constraint (24), while Section 4.3.2 deals with the case when the constraint is included.

4.3.1. Without equal-access constraint

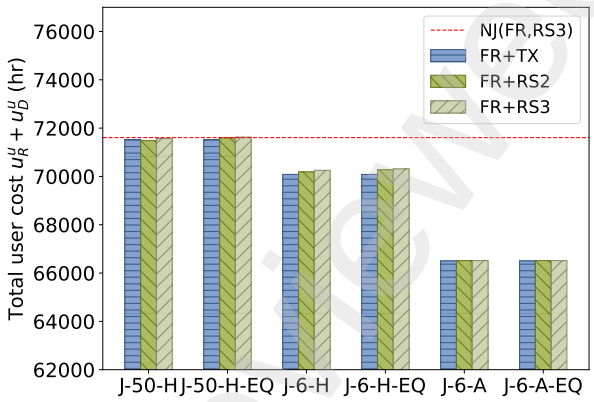
Figure 7 and Figure 8 illustrate how joint design improves, respectively, the total user cost and mode-specific average user cost, across different settings. First, it is important to note that when services are designed independently, neither budget level (baseline or 1.5 times baseline) is sufficient for TX. In contrast, while RS2 operates only at the higher budget level, RS3 functions at both levels; see Table B.3 in Appendix B for details. The total user cost reported in the plot (the horizontal red line) represents the sum of the FR user cost and the PT user cost under RS3. We choose RS3 as a representative of all PT modes because it is operable under most budget levels and it grants almost minimum total user costs with high budget levels under non-joint optimization in the status quo (according to Figure 6). It is also worth noting that in the status quo—and thus in the independent designs—the share of the PT budget is fixed at 12.2% (see Table B.3).

At first glance, the benefits of joint design appear negligible in J-50-H, the most comparable setting. Only in the higher budget scenario do we detect a slight improvement in the total user cost achieved by FR+RS2 and, to an even lesser extent, by FR+TX (see Figure 7). However, focusing solely on cost obscures a significant advantage of joint design: the ability to allocate the budget between the two services effectively, ensuring both operate properly. Indeed, while FR+TX does not function at either budget level under independent design due to insufficient budget for taxi services, it becomes viable when jointly designed. Table B.3 indicates that the joint design model allocates a higher share of the budget to taxi—nearly 20% of the baseline budget and 12.9% of the higher budget, compared to about 12.2% in the status quo.

In J-6-H with the baseline budget, the total user costs under joint design actually increase—regardless of the PT mode adopted—compared to the independent design (which is viable only for RS3). Indeed, as shown in Figure 7a, mode-specific user costs are slightly higher in J-6-H than in J-50-H. This occurs because, at this budget level, operating the FR service with six-seat human-driven pods is sub-optimal. Under a stringent budget, the operator is better off using larger vehicles that can meet demand with longer headways, thereby



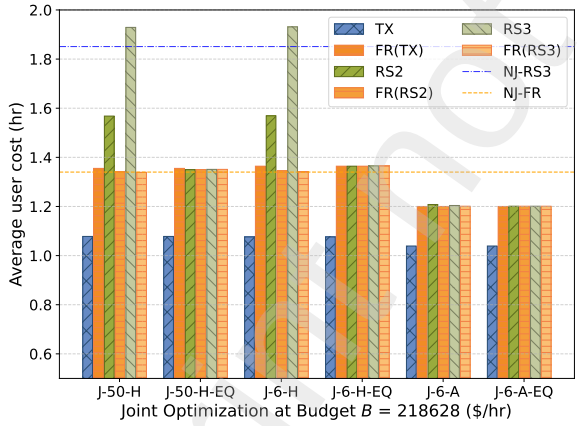
(a) Baseline budget ($B = 218638\$/hr$).



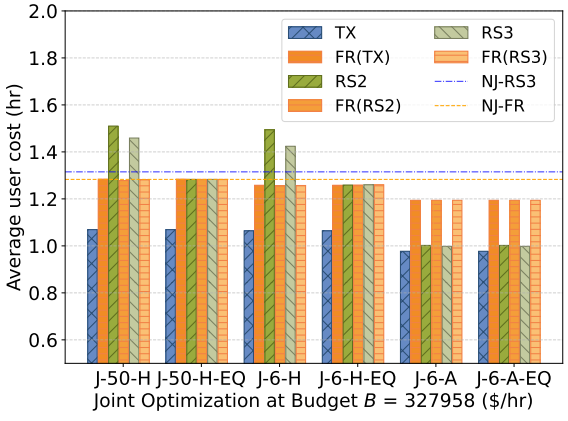
(b) 1.5 times baseline budget ($B = 327958\$/hr$).

Figure 7: User costs of three joint designs (FR+TX, FR+RS2, FR+RS3) vs. corresponding independent designs. J-50-H, J-6-H and J-6-A stand for 50-seat-pod + human-driven, six-seat-pod + human-driven, six-seat-pod + autonomous, respectively. Each setting has two versions: without equal-access constraint or with it (identified by -EQ at the end).

saving on labor costs. Being forced to use smaller pods results in low operational efficiency in the FR service, which ultimately drags down the performance of the joint design. However, Figure 7b suggests that under the higher budget, J-6-H outperforms both J-50-H and the status quo by a wide margin (approximately a 2% gain in total user cost). Figure 8 further indicates that this overall improvement is mostly attributable to the enhanced level of service enjoyed by FR users, while the average cost for PT users increases sharply compared with NJ, except in the case of the taxi mode.



(a) Baseline budget ($B = 218638\$/hr$).



(b) 1.5 times baseline budget ($B = 327958\$/hr$).

Figure 8: Mode-specific average user cost of three joint designs (FR+TX, FR+RS2, FR+RS3) vs. corresponding independent designs. J-50-H, J-6-H and J-6-A stand for 50-seat-pod + human-driven, six-seat-pod + human-driven, six-seat-pod + autonomous, respectively. Each setting has two versions: without equal-access constraint or with it (identified by -EQ at the end).

With six-seat pods and automation (J-6-A), the total user cost decreases by more than 10% under the baseline budget (see Figure 7). Additional improvements are observed with the higher budget, though they

are insignificant. An interesting finding is that, at the baseline budget, taxi outperforms the other PT modes in terms of total user cost—a competitive edge that disappears under the higher budget. As shown in Figure 8, while taxi users consistently experience a significantly better level of service than FR users under joint design at both budget levels, ridesharing users enjoy this advantage only at the higher budget level.

4.3.2. With equal-access constraint

From Figure 8 we can see that the joint design model with an equal-access constraint effectively enforces the requirement that PT users should not experience longer travel times than FR users. In both J-50-H and J-6-H, this constraint is clearly activated at both budget levels for PT modes `RS2` and `RS3` (as suggested by the identical average user cost for both services under those scenarios). For J-6-A, it is active only under the baseline budget, as the cost for PT users is substantially lower than that for FR users with the higher budget. Also worth noting is that the joint design operating the `TX` mode always renders a better service for PT users so that the equal-access constraint was never active.

Figure 7 highlights the negative impact of the equal-access constraint on total user cost. Although the constraint generally increases total user cost, the effect is small in most cases. The largest relative increase is observed for J-6-H with PT modes `RS2` and `RS3` under the baseline budget. As shown in Figure 8 and Table B.3, in these cases, imposing the constraint necessitates diverting a substantial portion of the budget from FR service to PT service, leading to a sharp rise in total user cost. For J-6-H without the constraint, the PT budget share is 13.09% and 11.67% for `RS2` and `RS3`, respectively; with the constraint, these shares increase to 17.3% and 16.83%. Even so, FR users experience an average cost increase of less than 2%, while PT users benefit from a relative gain of about 12%. Thus, although the equal-access constraint warrants consideration under tight budget conditions, its overall impact is expected to be modest.

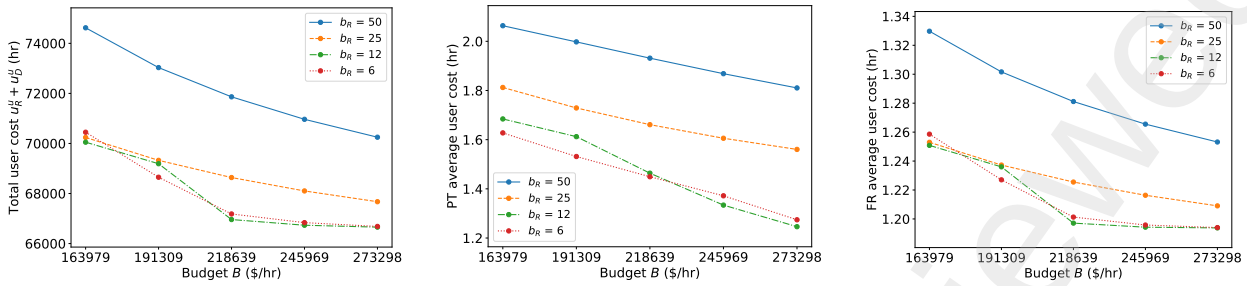
4.4. Sensitivity Analysis

The experiments conducted in this section are designed to test the sensitivity of joint design results to key parameters. We will focus on a design that combines FR with `RS3` mode. All pods are autonomous and, unless otherwise specified, have a seat capacity of six. In addition, the budget varies from 0.75 times the baseline value to 1.25 times that amount. We consider three parameters: seat capacity (Section 4.4.1), demand rate (Section 4.4.2) and pick-up/drop-off time in PT service (Section 4.4.3).

4.4.1. Seat capacity

Figure 9 reports the three main performance metrics (the total user cost and the average cost of PT and FR users) of the joint design under different budgets and seat capacities.

The joint design tends to prefer smaller pods to larger ones, though the difference is substantial only for the very large seat capacity (50). The pattern is largely consistent with what we have seen in Figure 4, which reports the results of an independent FR design. Interestingly, even though seat capacity only



(a) Total user cost. (b) Average PT user cost. (c) Average FR user cost.

Figure 9: Sensitivity of joint design results to seat capacity b_R. Autonomous pods operating FR+RS3.

changes in FR service, it has a much bigger impact on PT service. For example, at the baseline budget, when seat capacity is reduced from 50 to 6, FR users experience a cost reduction of about 7%, compared to above 20% for PT users.

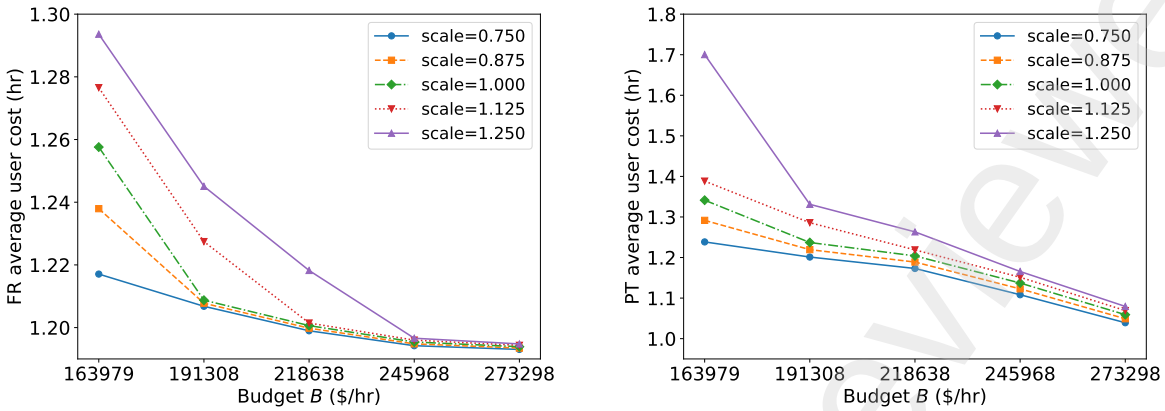
When the budget is tight, the joint design favors FR service over PT service regardless of seat capacity. At 0.75 times the baseline budget and using six-seat pods, PT users experience an average travel time of about 1.65 hours—more than 30% higher than that of the average FR user. As the budget increases, the design begins allocating more resources to FR service. Near parity is reached at the 1.25 times baseline budget when FR service is operated with smaller pods (6 or 12 seats). However, for seat capacities of 25 and 50, FR users remain better off. Because smaller pods improve efficiency, the design more quickly reaches the threshold of diminishing returns on investment for FR service—beyond that point, resources begin to shift toward PT service, where they can be used more productively.

4.4.2. Demand

We perform two experiments. In the first, a uniform scalar, ranging between 0.75 and 1.25, is applied to both FR and PT demand. The second experiment holds the total demand constant, while allowing the ratio of PT demand to vary from 1% (approximately the level at the status quo) to 5%. Figures 10 and 11 report the results, respectively, for the first and the second experiment.

In the first experiment, as expected, a higher demand tends to increase the average cost for both PT and FR users when the budget is held constant, while increasing the budget has the opposite effect (see Figure 10). As the budget increases, the impact of the demand level becomes less pronounced. Moreover, PT users are more affected by increasing demand at a given budget than FR users. For example, under the baseline budget, increasing the uniform demand scalar from 0.75 to 1.25 results in a relative cost increase that is approximately four times higher for PT users compared to FR users. This gap grows wider when the budget is tighter.

For the second experiment, Figure 11 demonstrates that, as the ratio of PT demand increases, the FR service level degrades. The impact is more pronounced under tighter budgets. For example, Figure 11a

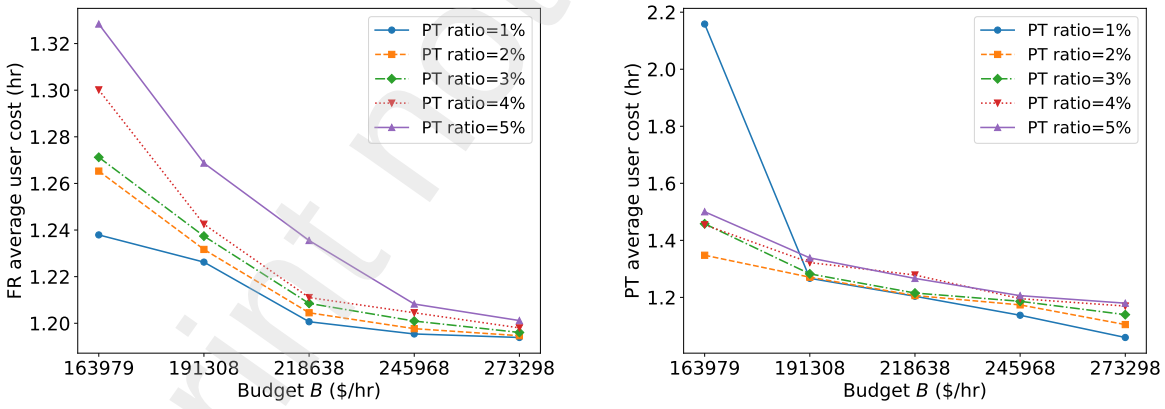


(a) Average cost of FR users.

(b) Average cost of PT users.

Figure 10: Sensitivity of joint design results to a uniform demand scalar applied to the base FR and PT demands. Six-seat autonomous pods operating FR+RS3.

shows that at 0.75 times the baseline budget, FR users experience an approximate 3.6-minute increase in the average cost (about 5%) when the ratio of PT users rises from 1% to 5%. A similar trend is observed for PT service (Figure 11b), where a higher PT ratio tends to worsen the average cost for its users, although the effect is not as clear-cut[7]. Thus, serving a greater number of PT users exerts downward pressure on the overall level of service—a result that is hardly surprising, given that providing PT services is financially more demanding on a per-capita basis.



(a) Average cost of FR users.

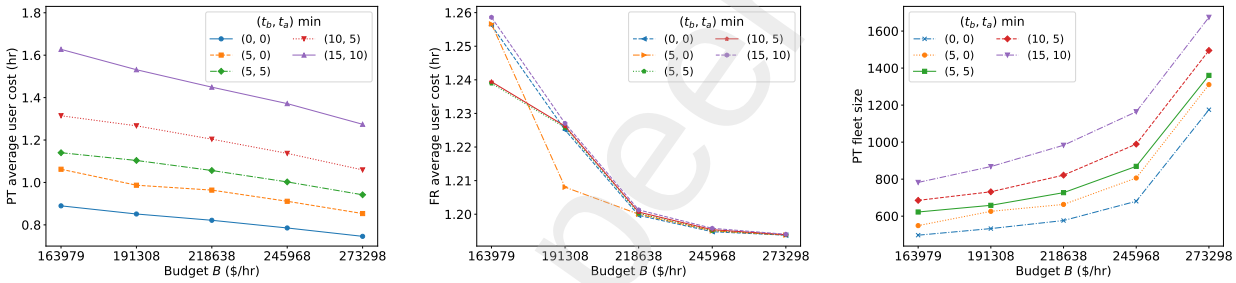
(b) Average cost of PT users.

Figure 11: Sensitivity of joint design results to different PT demand ratio (the total demand is fixed at the base level). Six-seat autonomous pods operating FR+RS3.

[7]At 0.75 times the baseline budget, there is a noticeable jump in PT user cost that appears anomalous. A closer examination reveals that this anomaly is caused by the integral requirement for the number of p-trains used in the FR service: a lower budget increases the returns on investment for the FR service, thereby allocating a higher percentage of the dwindling budget to FR service (so that a smaller p-train may be used) and, consequently, disproportionately affecting PT users.

4.4.3. Boarding and alighting time

Figure 12 illustrates how the performance of the joint design varies with the boarding and alighting times required by PT service. The results indicate that PT user costs are significantly affected by these operations. For example, Figure 12a shows that at the baseline budget, the default boarding/alighting times ($t_b = 10$, $t_a = 5$) increase the expected PT user cost by approximately 0.4 hours (or 50%) compared to the scenario where these times are set to zero. This increase occurs despite a corresponding rise in fleet size—from about 550 when $t_b = t_a = 0$ to over 800 (see Figure 12c) at the default times—indicating that the extra pods are needed to cover the waiting associated with boarding and alighting rather than to improve service levels. In contrast, Figure 12b shows that the effect on FR users is minimal, with the worst-case increase in average cost being less than 2%. These findings underscore the importance of efficient boarding and alighting operations for PT services and highlight the limited capability of resource allocation to mitigate the negative impact of inefficiencies in these operations.



(a) Average cost of PT users. (b) Average cost of FR users. (c) PT fleet size.

Figure 12: Sensitivity of joint design results to boarding and alighting time (t_b, t_a) in the PT service. Six-seat autonomous pods operating FR+RS3.

5. Conclusion

We examined how newly emerged modular autonomous vehicles (pods) could enhance the overall efficiency of transit operations. Pods offer several promising advantages. First, they operate without drivers, which could significantly reduce labor costs. Second, they are likely powered by electricity—a cleaner and more cost-effective alternative to fossil fuels. Finally, the modularity of pods enables more agile operations, making it feasible to run both paratransit (PT) and regular fixed-route (FR) services with a single fleet. Our investigation is motivated by this possibility and focuses on the trade-offs that transit operators must consider when contemplating this futuristic scenario.

Our case study, based on transit data from the Chicago region, led to the following findings:

Modularity. Modularity enables the formation of pod trains with small pods, which benefits FR operations. Our experiments recorded a maximum average user cost savings of a few percent attributable to

modularity. When the budget is abundant and labor costs are low, smaller pods tend to provide greater benefits.

Automation. Everything else being equal, automation delivers greater service improvements to PT users than to FR users. This discrepancy arises from the fact that PT service has a more labor-intensive cost structure on a per-capita basis than FR service. Furthermore, the benefits of automation are generally more pronounced when the budget is tighter.

Joint design Jointly designing the two services helps avoid resource misallocation, which in the worst cases could render a service dysfunctional due to an insufficient budget. However, without leveraging modularity and automation, the potential for cost reduction through joint design is limited. When the budget is tight, imposing the equal-access constraint tends to improve service for PT users at the expense of FR users. On a per-capita basis, though, the extra cost incurred by FR users is much smaller than the savings enjoyed by PT users. Thus, enforcing the equal-access clause is unlikely to provoke stiff opposition.

PT modes. Ridesharing is operationally flexible, allowing it to offer a wider range of service levels based on the available budget. Within ridesharing modes, greater operational flexibility tends to improve efficiency at the expense of service quality.

While the present study offers a novel approach to addressing the challenge of rapidly rising demand for paratransit services, it has several limitations. We briefly discuss three of these in the following, and suggest directions for future research accordingly.

First, paratransit is modeled here as a steady queuing system similar to an on-demand ride-hail service. In reality, paratransit may possess unique attributes, such as high reservation rates and special requirements (e.g., specific equipment or human assistance). Future research should examine whether and how these features can be incorporated into strategic transit design.

Second, an important advantage of modularity is its ability to manage within-day variations in transit demand. This benefit becomes even more prominent if PT demand can be shifted to avoid peak-period competition with FR demand. To quantify such benefits, however, a model that endogenizes the allocation of resources and paratransit demand across multiple periods is needed. Developing such a model represents another valuable extension of the present work.

Finally, with autonomous pods, FR and PT services could be more tightly integrated at the operational level. For example, a pod might pick up a few PT riders before joining a pod train serving an FR line, with riders then transferring to other pods to reach their destinations. Modeling the impact of such operational flexibilities on overall joint design performance presents another challenge that we leave to future studies.

6. Acknowledgments

This research is funded by Center for Connected and Automated Transportation (CCAT), part of the University Transportation Center Program sponsored by US Department of Transportation (USDOT).

References

Chen, P. W., & Nie, Y. M. (2017). Analysis of an idealized system of demand adaptive paired-line hybrid transit. *Transportation Research Part B: Methodological*, *102*, 38–54. URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261516305562. doi:10.1016/j.trb.2017.05.004.

Chen, Z., & Li, X. (2021). Designing corridor systems with modular autonomous vehicles enabling station-wise docking: Discrete modeling method. *Transportation Research Part E: Logistics and Transportation Review*, *152*, 102388. doi:10.1016/j.tre.2021.102388.

Chen, Z., Li, X., & Qu, X. (2021). A Continuous Model for Designing Corridor Systems with Modular Autonomous Vehicles Enabling Station-wise Docking. *Transportation Science*, . doi:10.1287/trsc.2021.1085.

Cheng, X., Nie, Y. M., & Lin, J. (2024). An autonomous modular public transit service. *Transportation Research Part C: Emerging Technologies*, (p. 104746).

Chicago Transit Authority (2020a). *2019 Annual Ridership Report*. Technical Report Chicago Transit Authority. URL: https://www.transitchicago.com/assets/1/6/2019_Annual_Ridership_Report.pdf accessed: 2025-02-12.

Chicago Transit Authority (2020b). Facts at a glance. URL: https://web.archive.org/web/20200226193605/https://www.transitchicago.com/facts/ accessed: February 13, 2025.

Daganzo, C. F. (1978). An approximate analytic model of many-to-many demand responsive transportation systems. *Transportation Research*, *12*, 325–333. URL: http://www.sciencedirect.com/science/article/pii/0041164778900072. doi:https://doi.org/10.1016/0041-1647(78)90007-2.

Daganzo, C. F. (2010). Structure of competitive transit networks. *Transportation Research Part B: Methodological*, *44*, 434–446.

Daganzo, C. F., & Ouyang, Y. (2019a). A general model of demand-responsive transportation services: From taxi to ridesharing to dial-a-ride. *Transportation Research Part B: Methodological*, *126*, 213–224. URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261518307793. doi:10.1016/j.trb.2019.06.001.

Daganzo, C. F., & Ouyang, Y. (2019b). *Public transportation systems: Principles of system design, operations planning and real-time control*. World Scientific.

Dai, T., Zheng, H., & Nie, M. (2024). Is fare free transit just? quantifying the impact of moral principles on transit design and finance. *Quantifying the Impact of Moral Principles on Transit Design and Finance (May 24, 2024)*, .

Federal Transit Administration (2022). The National Transit Database. https://www.transit.dot.gov/ntd.

Fu, L., & Ishkhanov, G. (2004). Fleet Size and Mix Optimization for Paratransit Services. *Transportation Research Record: Journal of the Transportation Research Board*, *1884*, 39–46. URL: http://journals.sagepub.com/doi/10.3141/1884-05. doi:10.3141/1884-05.

Guo, R., Zhang, W., Guan, W., & Ran, B. (2021). Time-Dependent Urban Customized Bus Routing With Path Flexibility. *IEEE Transactions on Intelligent Transportation Systems*, *22*, 2381–2390. URL: https://ieeexplore.ieee.org/abstract/document/9186321. doi:10.1109/TITS.2020.3019373. Conference Name: IEEE Transactions on Intelligent Transportation Systems.

Gupta, D., Chen, H.-W., Miller, L. A., & Surya, F. (2010). Improving the efficiency of demand-responsive paratransit services. *Transportation research part A: policy and practice*, *44*, 201–217.

Jaffe, E. (2015). Uber and public transit are trying to get along. *Citylab*, . URL: http://www.citylab.com/cityfixer/2015/08/uber-and-public-transit-are-trying-to-get-along/400283/.

Jara-Díaz, S., Fielbaum, A., & Gschwender, A. (2020). Strategies for transit fleet design considering peak and off-peak periods using the single-line model. *Transportation Research Part B: Methodological*, *142*, 1–18. URL: https://www.sciencedirect.com/science/article/pii/S0191261520304045. doi:10.1016/j.trb.2020.09.012.

Kane, J., Tomer, A., & Puentes, R. (2016). *How Lyft and Uber can improve transit agency budgets.*. Technical Report The Brookings Institution.

Kaufman, S. M., Smith, A., O'Connell, J., & Marulli, D. (2016). *Intelligent Paratransit*. Technical Report New York University, Rudin Center for Transportation Research and Policy.

Khan, Z. S., He, W., & Menéndez, M. (2023). Application of modular vehicle technology to mitigate bus bunching. *Transportation Research Part C: Emerging Technologies*, *146*, 103953.

Khan, Z. S., & Menéndez, M. (2023). Bus splitting and bus holding: A new strategy using autonomous modular buses for preventing bus bunching. *Transportation Research Part A: Policy and Practice*, *177*, 103825.

Khan, Z. S., & Menendez, M. (2024). A seamless bus network without external transfers using autonomous modular vehicles. *Available at SSRN 4899526*, .

Lin, J., Nie, Y. M., & Kawamura, K. (2023). An Autonomous Modular Mobility Paradigm. *IEEE Intelligent Transportation Systems Magazine*, *15*, 378–386. URL: https://ieeexplore.ieee.org/document/9750205/. doi:10.1109/MITS.2022.3159484.

Liu, H., Zhao, Y., Li, J., Li, Y., & Gao, X. (2022). Optimization Model of Transit Route Fleet Size Considering Multi Vehicle Type. *Sustainability*, *14*, 193. URL: https://www.mdpi.com/2071-1050/14/1/193. doi:10.3390/su14010193. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Liu, T., Ceder, A., & Rau, A. (2020). Using Deficit Function to Determine the Minimum Fleet Size of an Autonomous Modular Public Transit System. *Transportation Research Record*, *2674*, 532–541. doi:10.1177/0361198120945981.

Liu, X., Qu, X., & Ma, X. (2021). Improving flex-route transit services with modular autonomous vehicles. *Transportation Research Part E: Logistics and Transportation Review*, *149*, 102331. doi:10.1016/j.tre.2021.102331.

Liu, Y., Chen, Z., & Wang, X. (2024). Alleviating bus bunching via modular vehicles. *25th Internat. Sympos. Transportation Traffic Theory ISTTT25 (Ann Arbor, MI)*, .

Miah, M. M., Naz, F., Hyun, K. K., Mattingly, S. P., Cronley, C., & Fields, N. (2020). Barriers and opportunities for paratransit users to adopt on-demand micro transit. *Research in transportation economics*, *84*, 101001.

Nguyen-Hoang, P., & Yeung, R. (2010). What is paratransit worth? *Transportation Research Part A: Policy and Practice*, *44*, 841–853. URL: https://linkinghub.elsevier.com/retrieve/pii/S0965856410001278. doi:10.1016/j.tra.2010.08.006.

Ouyang, Y., Nourbakhsh, S. M., & Cassidy, M. J. (2014). Continuum approximation approach to bus network design under spatially heterogeneous demand. *Transportation Research Part B: Methodological*, *68*, 333–344. URL: https://linkinghub.elsevier.com/retrieve/pii/S0191261514001076. doi:10.1016/j.trb.2014.05.018.

Pace Suburban Bus Service (2020a). 2019 final budget. URL: https://www.pacebus.com/sites/default/files/2020-06/2019_Final_Budget.pdf.

Pace Suburban Bus Service (2020b). 2020 annual financial report. URL: https://www.pacebus.com/sites/default/files/2020-06/2019%20Annual%20Financial%20Report.pdf.

Pei, M., Lin, P., Du, J., Li, X., & Chen, Z. (2021). Vehicle dispatching in modular transit networks: A mixed-integer nonlinear programming model. *Transportation Research Part E: Logistics and Transportation Review*, *147*, 102240. doi:10.1016/j.tre.2021.102240.

Rahimi, M., Amirgholy, M., & Gonzales, E. J. (2014). *Continuum Approximation Modeling of ADA Paratransit Operations in New Jersey*. Technical Report.

Regional Transportation Authority Mapping and Statistics (2024). Pace vanpool ridership data. https://rtams.org/

ridership/pace/vanridership. Accessed: 2024-07-31.

Rodriguez-Roman, D., López-Martínez, J., & Figueroa-Medina, A. M. (2024). c bus networks and complementary paratransit service areas. *Transportation Letters*, (pp. 1–13). URL: https://www.tandfonline.com/doi/full/10.1080/19427867.2024.2320499. doi:10.1080/19427867.2024.2320499.

Stein, D. M. (1978). Scheduling dial-a-ride transportation systems. *Transportation Science*, *12*, 232–249.

Tian, Q., Lin, Y. H., Wang, D. Z., & Liu, Y. (2022). Planning for modular-vehicle transit service system: Model formulation and solution methods. *Transportation Research Part C: Emerging Technologies*, *138*, 103627.

United States Government Accountability Office (2012). *ADA PARATRANSIT SERVICES: Demand Has Increased, but Little Is Known about Compliance.*. Technical Report United States Government Accountability Office.

University of Chicago (2024). The grid. URL: https://chicagostudies.uchicago.edu/grid accessed on 2024-12-28.

Wilson, N. H., Weissberg, R. W., & Hauser, J. (1976). *Advanced Dial-a-Ride Algorithms Research Project*. Technical Report.

Wu, J., Kulcsár, B., Selpi, & Qu, X. (2021). A modular, adaptive, and autonomous transit system (MAATS): A in-motion transfer strategy and performance evaluation in urban grid transit networks. *Transportation Research Part A: Policy and Practice*, *151*, 81–98. doi:10.1016/j.tra.2021.07.005.

Zhang, Z., Tafreshian, A., & Masoud, N. (2020). Modular transit: Using autonomy and modularity to improve performance in public transportation. *Transportation Research Part E: Logistics and Transportation Review*, *141*, 102033.

Zheng, H., Li, J., Lin, J., & Nie, M. (2024). Exploiting modularity for co-modal passenger-freight transportation. *Available at SSRN 4963867*, .

Appendix A. Estimation of model inputs

Appendix A.1. Demand, budget and fleet size

To estimate the hourly demand, we assume both FR and PT services operate 16 hours a day, and choose a peak hour factor (PHF_{fr}) of 0.085 for FR (Dai et al., 2024) and 0.0625 for PT (i.e., no peaking effect for PT). Thus, for FR, the peak hour demand is estimated as $237,300,000 \div 365 \times 0.085 \div 803 = 68.8$ pax/hr/km^2 and the peak hour budget is $824,288,048 \div 365 \times 0.085 = \$191,957$. Recalling that the PT demand falling into the CTA service area (the area for the case study) is about 76% of all Pace Paratransit rides, we estimate the PT demand rate at $4,280,000 \times 0.76 \div 365 \div 16 \div 803 = 0.691$ pax/hr/km^2.

We further estimate about 83.5% of the Paratransit budget was spent on the rides originated within the CTA service area, leading to a total Paratransit budget of about \$155.6 million, or about \$26,681 per hour. To see why, note that, according the Pace budget in 2019 (Pace Suburban Bus Service, 2020b), the total private contractor cost for Paratransit service was about \$163 million, of which about \$136 million were for the city of Chicago. This means (i) about \$186 - \$163 = \$23 million were non-contract spending on Paratransit (ii) the Paratransit cost in Chicago roughly accounts for about 83.5% of the Pace's Paratransit budget. We thus estimate the total Paratransit budget for the city of Chicago, inclusive of both contract and non-contract costs, as \$136 +\$23×0.835 ∼ \$156 million. Using the same ratio, we estimate the number of vehicles employed for our case study is $552 \times 0.835 = 461$.

Appendix A.2. Calibration

To choose $\overline{c}_m$, $\overline{c}_D$ and $\overline{c}_M$, we ensure the current FR service, as specified in our model, spends \$191,957 per peak hour, while employing a fleet of 1,861 50-seat buses to serve a demand 68.8 pax/hr/km^2. We estimate the number of lines N_0 in the status quo as 70, obtained based on $28/0.4 = 70$, where $0.4km$ is the average block length in Chicago (University of Chicago, 2024). The status quo headway, H_0, is calibrated along with $\overline{c}_D$ and $\overline{c}_M$. We found that when $H = 12.5$ min, $\overline{c}_m = \$9$/veh, $\overline{c}_D = \$0.8$/km and $\overline{c}_M = \$38$/p-train in a revenue hour, the design consumes nearly all the budget for the given fleet size, see Table 2. To choose$\underline{c}_m$, $\underline{c}_D$, and $\underline{c}_M$, we match the ridership, fleet size and budget of the current PT service. We assume that the current PT service operates 461 six-seat vans in the `DR` mode with $b_D = 3$. We found that when $\underline{c}_m$ =\$1.5/km $\underline{c}_D$ =\$0.4/km and $\underline{c}_M$ =\$9/p-train, the PT services consumes nearly all the budget for the given fleet size at the steady state (with the number of waiting passengers $z = 18$), see Table 2.

We assume that, under the status quo, the FR service operates with a 50-seat bus, while the PT service uses a six-seat van. The calibrated acquisition cost, distance cost, and non-driver time cost for the FR service are approximately six times, two times, and five times higher, respectively, than those for the PT service. These cost differences fall within a reasonable range.

Appendix B. Joint design results

Table B.3: Average user travel time of joint designs vs. independent designs. NA means the result is not available due to insufficient budget; AC stands for the share of PT budget in the total budget.

Budget	Scenarios	FR+TX			FR+RS2			FR+RS3		
		TX (hr)	**FR (hr)**	**AC (%)**	**RS2 (hr)**	**FR (hr)**	**AC (%)**	**RS3 (hr)**	**FR (hr)**	**AC (%)**
218638	**NJ-50-H**	NA	**1.340**	**12.20**	NA	**1.340**	**12.20**	**1.851**	**1.340**	**12.20**
	J-50-H	**1.078**	**1.355**	**19.29**	**1.568**	**1.342**	**13.09**	**1.929**	**1.339**	**11.68**
	J-50-H-EQ	1.078	1.355	19.29	1.350	1.350	17.04	1.351	1.351	17.56
	J-6-H	**1.076**	**1.364**	**17.01**	**1.569**	**1.345**	**13.09**	**1.931**	**1.342**	**11.67**
	J-6-H-EQ	1.076	1.364	17.01	1.364	1.364	17.30	1.365	1.365	16.83
	J-6-A	**1.039**	**1.200**	**5.62**	**1.208**	**1.200**	**6.76**	**1.204**	**1.201**	**7.07**
	J-6-A-EQ	1.039	1.200	5.62	1.201	1.201	6.89	1.201	1.201	7.12
327958	**NJ-50-H**	NA	**1.283**	**12.20**	**1.288**	**1.283**	**12.20**	**1.315**	**1.283**	**12.20**
	J-50-H	**1.069**	**1.284**	**12.90**	**1.510**	**1.278**	**9.05**	**1.459**	**1.281**	**10.65**
	J-50-H-EQ	1.069	1.284	12.90	1.283	1.283	12.29	1.283	1.283	12.75
	J-6-H	**1.064**	**1.258**	**11.40**	**1.494**	**1.255**	**9.18**	**1.424**	**1.257**	**10.93**
	J-6-H-EQ	1.064	1.258	11.41	1.259	1.259	12.75	1.260	1.260	13.22
	J-6-A	**0.977**	**1.194**	**23.18**	**1.002**	**1.194**	**23.41**	**0.998**	**1.194**	**23.38**
	J-6-A-EQ	0.977	1.194	23.18	0.998	1.194	23.38	0.998	1.194	23.38