Vehicle Routing for Shared-Mobility Systems with Time-Varying Demand

Kevin Spieser, Samitha Samaranayake, and Emilio Frazzoli

Abstract—This work considers mobility systems in which a shared fleet of self-driving vehicles is used to transport passengers. More specifically, we focus on policies to route both passenger-filled and empty vehicles when the travel demand is time-varying. In this setting, we argue that metrics, such as the cost to relocate empty vehicles, which are well-defined in a stand-alone capacity under steady-state conditions, now make sense only within a framework that reflects inherent tradeoffs with other metrics, e.g., the fleet size and the quality of service provided. As a first step toward developing a general theory of time-varying, shared-mobility systems, we provide an optimization framework that models passengers and vehicles as continuous fluids, and their movement as fluid flows. The model is used to develop some initial performance results related to the minimum number of vehicles required to avoid passenger queueing. Finally, simulation results of a hypothetical shared mobility system based in Singapore demonstrate how a fleet manager could use our optimization approach to select a vehicle routing policy.

I. INTRODUCTION

Utopian visions of future transportation systems often feature a large user base that moves about an urban environment by sharing access to a fleet of automated vehicles. In this setting, passengers are picked up at their current location, driven to their destination, and dropped off. When a vehicle becomes empty, its autonomous capabilities allow it to drive off and serve other passengers. We refer to systems with this functionality as Automated Mobility-on-Demand or AMoD systems. Compared to owning a private vehicle, AMoD systems feature higher vehicle utilization rates and offer the financial savings of a shared economy [1], [2].

Owing to suburbanization and entrenched conventions, such as starting work in the early morning, urban travel patterns almost always display a time-varying profile. Often the temporal fluctuations in terms of both the intensity and spatial preference for travel are pronounced. For example, most large cities experience a major influx of individuals during the morning commute to work. In the evening, the same cities witness a comparable exodus of people when workers return home. To ensure fleet vehicles do not sit idle for extended periods of time, the supply of empty vehicles must, on some timescale, be realigned with the demand for travel. The process of re-locating empty vehicles from popular dropoff locations to popular pickup locations is referred to as rebalancing [3]. In the context of an AMoD system, an operational policy must therefore specify not only (i) how customers should be routed between origin and destination points, but also (ii) how empty vehicles should be rebalanced. Given a fixed network and a known transportation demand, the policy directly determines important metrics, including the fleet size, the rebalancing cost, and the average number of passengers waiting, in a queue, for a vehicle. Unfortunately, it is generally impossible to optimize one of these quantities without adversely affecting another. For example, minimizing queueing effects generally requires fielding additional fleet vehicles, a more aggressive approach to rebalancing, or both.

Attempts to rigorously study AMoD systems generally focus on policies that are optimal with respect to one of the aforementioned metrics. Moreover, to the best of our knowledge, all of the theoretically-oriented rebalancing results assume a stationary demand model, i.e., the intensity and spatial preference for travel are constant over the period of interest. The primary contribution of this paper is to begin the process of understanding the relationship among fleet size, rebalancing cost, and queueing effects in an AMoD system imposed by a time-varying travel demand. In this direction, we illustrate how a fleet operator could use the ideas presented in this paper to select a routing policy for a hypothetical AMoD system using historical travel data. Long term, the hope is that our results support the development of a general theory of time-varying AMoD systems.

The remainder of the paper is organized in sections. Section II reviews shared-vehicle systems and existing approaches to rebalance vehicle fleets. Section III defines the Earth Mover’s Distance between two probability distributions and Pareto optimal policies. These ideas will feature prominently in the analysis that follows. The notation and terminology used to describe the components and performance of AMoD systems is listed in Section IV. A fluid model describing the operation of an AMoD system with time-varying demand is given in Section V. Section VI shows that the model reduces to a previously reported result in the case of stationary demand. In Section VII, we show that one form of worst case performance is realized for a disjoint demand model. Simulation results are described in Section VIII for a hypothetical AMoD system based in Singapore. Finally, Section IX closes by summarizing the key ideas and describing a subset of future work items.
II. RELATED WORK

AMoD systems represent the fusion of autonomous vehicle technology and one-way car sharing systems. Multiple organizations are now heavily invested in autonomous vehicles and advances have been widely reported [4]. For this paper, it suffices to know that an autonomous vehicle can drive itself from one location to another. Accordingly, we do not delve into the advancements in the sensing, planning, and control strategies that make this possible.

To support our interest in operations at the fleet level, we devote this section to providing a brief history of carsharing systems, followed by an overview of strategies to rebalance empty vehicles.

The earliest carsharing systems originated in Europe during the mid-twentieth century [5]. Their emergence in North America has been much more recent. Currently, the majority of carsharing services are two-way, meaning a vehicle must be dropped off at the same location it was originally picked up from. To attract a larger clientele, one-way carsharing systems have emerged [6], [7]. Here, vehicles may be dropped off at a destination of the user’s choice. Well-designed, one-way systems offer users greater freedom, but require rebalancing to prevent vehicles from piling up at popular dropoff locations and being in short supply at popular pickup locations.

Researchers have noted that demand for one-way services is closely tied to the availability of vehicles [5]. This awareness has spurred interest in the closely connected issues of fleet sizing and rebalancing [8], [9], [10], [11], [12]. Early approaches to understand rebalancing often made unrealistic assumptions, e.g., that the pickup and dropoff distributions are equal, and relied heavily on simulation. When provided with a specific rebalancing policy, these tools are capable of quantifying the distribution and availability of vehicles. Unfortunately, using these tools for rebalancing often involves relying on heuristics or tweaking only a subset of the decision variables, and, therefore, offers no guarantee of optimality.

More recently, [13] considers the problem of coordinating rebalancing efforts using a fluid model to describe the flows of passenger-filled and empty vehicles between stations in a network. For steady-state conditions, the fleet of minimum size that provides stability is found by solving a linear program. The policy requires perfect knowledge of the transportation demand, i.e., the rate at which users arrive at station $i$ heading to station $j$, to enforce the conservation of vehicles at all stations.

In [14], the theory of Jackson networks is used to cast the rebalancing problem in a queuing framework. Vehicles are rebalanced so that the probability of an arriving customer finding a vehicle is equal across stations. However, no customer queues form in their model because customers are assumed to be impatient. That is, if a customer arrives at a station devoid of vehicles, they immediately leave the system. Theoretical guarantees again require knowledge of the travel demand, but a feedback policy that does not and periodically redistributes empty vehicles evenly across all stations is also provided.

While the aforementioned works provide linear programs that guarantee optimality with respect to one aspect of performance, they assume static demand profiles and focus on a single measure of performance. In this work, we consider vehicle routing strategies for AMoD systems with time-varying demands, and argue that the correct interpretation is to consider the tradeoffs between various performance measures. In this way, the work is similar in spirit, though different in venue, to [15], [16], and [17], which consider performance tradeoffs between delay, throughput, and energy consumption for a variety of communication networks.

III. MATHEMATICAL PRELIMINARIES

This section reviews two ideas that feature prominently in later sections: (i) the Earth Mover’s Distance, a measure of distance between two distributions, and (ii) Pareto optimality, the notion that improvements in one aspect of performance may be realized only at the expense of another.

The Earth Mover’s Distance has been used as a metric in image processing, artificial intelligence, and, most aptly for our needs, transportation systems to measure the effort required to transform one distribution into another, see, for example, [18], [19], and [20]. Formally, given a set $X$ and two distributions $\varphi_1$ and $\varphi_2$, where $\varphi_i : X \rightarrow \mathbb{R}_{\geq 0}$ for $i = 1, 2$, and a distance metric $D$ on $X$, the Earth Mover’s Distance, denoted $\text{EMD}(\varphi_1, \varphi_2)$, is the first Wasserstein distance [21]. Mathematically,

$$\text{EMD}(\varphi_1, \varphi_2) = \inf_{\gamma \in \Gamma(\varphi_1, \varphi_2)} \int_X \int_X D(x_1, x_2)d\gamma(x_1, x_2),$$

where $\Gamma(\varphi_1, \varphi_2)$ is the set of all measures with marginals $\varphi_1$ and $\varphi_2$ on the first and second factor, respectively. Informally, if $\varphi_1$ and $\varphi_2$ represent two piles of “dirt” (i.e., earth), then $\text{EMD}(\varphi_1, \varphi_2)$ is the minimum work (dirt distance) required to reshape $\varphi_1$ into $\varphi_2$, or vice versa.

In taxi systems with time-invariant pickup and dropoff distributions $\varphi_0$ and $\varphi_d$, respectively, $\text{EMD}(\varphi_0, \varphi_d)$ is a lower bound on the minimum distance a vehicle must travel, on average, to transition from one job to the next [22]. Later, we argue that in the time-varying case, the cost of rebalancing vehicles only makes sense in relation to other performance metrics, e.g., fleet size and queuing effects. The following concept is useful when considering tradeoffs between these metrics for different policies.

Definition 3.1: (Pareto Optimality) Consider a system for which performance is characterized by $M$ attributes, $a_1, \ldots, a_M$. Let $A$ be the set of all realizable attribute tuples. With respect to attribute $i$, $i = 1, \ldots, M$, $a_i'' \succeq a_i'$ denotes that $a_i''$ is no worse than $a_i'$, and $a_i'' > a_i'$ denotes that $a_i''$ is strictly preferable to $a_i'$. The tuple $(a_1'', \ldots, a_M'')$ is Pareto optimal if $\forall(a_1'', \ldots, a_M'') \in A$, $a_i'' < a_i'$ for some $i = 1, \ldots, M$, implies $\exists j \neq i$ for which $a_j'' < a_j'$.
We will be interested exclusively in Pareto optimal routing policies. That is, while it may be possible, usually by doing something unreasonable, e.g., sending empty vehicles on unnecessary trips, to realize combinations of attributes that are not Pareto optimal, these policies do not speak to fundamental performance limitations because they contain inherent room for improvement.

IV. NOTATION AND TERMINOLOGY

This section develops the notation and terminology used to describe AMoD systems. It is useful to view an AMoD system in terms of its demand side and its supply side. The first two subsections reflect this dichotomy. The final subsection describes the performance metrics we will be interested in.

A. The Demand Side of an AMoD System

The demand side of an AMoD system describes the travel patterns of passengers and how they move through the system. Formally, the demand is given by a tuple

\[ \text{Demand} : (X, \lambda, \varphi, w). \]  

(2)

In (2), \( X \) is a finite set of stations. Passengers enter the system by arriving at a station and exit the system by departing from a station. These stations are referred to as the passenger’s origin and destination, respectively. Passengers travel from their origin to their destination by vehicle. Vehicles can travel directly from any station to any other station. We refer to a passenger’s request for transport as a demand and say that a demand has been served when the passenger is transported from their origin to their destination.

The scalar function \( \lambda : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0} \) describes the intensity of the demand, with \( \lambda(t) \) denoting the rate at which demands enter the system at time \( t \).

The probability distribution \( \varphi : X \times X \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0} \), and satisfying

\[ \sum_{i \in X} \sum_{j \in X} \varphi_{ij}(t) = 1, \]  

(3)

describes the shape of the travel demand as a function of time, with \( \int_{t_1}^{t_2} \lambda(t) \varphi_{ij}(t) dt \) the number of passengers that arrive at station \( i \), between times \( t_1 \) and \( t_2 \) with \( t_2 \geq t_1 \), traveling to station \( j \).

Finally, the function \( w : X \times X \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0} \), with \( \int_{0}^{\infty} w_{ij}(t, \tau) d\tau = 1 \) for all \( i, j \in X \), \( t \in \mathbb{R} \), specifies the distribution of travel times between stations, with \( \int_{t_1}^{t_2} w_{ij}(t, \tau) d\tau \) the probability of leaving \( i \) at time \( t \) and arriving at \( j \) between times \( t_1 \) and \( t_2 \). Clearly, \( w_{ij}(t_1, t_2) = 0 \) for \( t_2 < t_1 \). It is assumed that \( \varphi_{ij}(\cdot) \) and \( w_{ij}(\cdot, \cdot) \) are periodic with period \( T_P \), so that, \( \varphi_{ij}(t) = \varphi_{ij}(t + T_P) \), for all \( i, j \in X \), \( t \in \mathbb{R} \). In the case of \( T_P = 24 \) hours, this periodicity captures the daily cycles in travel patterns during weekdays. For convenience, define

\[ \varphi_{ij} = \frac{1}{T_P} \int_{0}^{T_P} \varphi_{ij}(t) dt \]  

(4)

to be the time-averaged spatial demand. Similarly, the associated origin and destination marginal density functions are given by

\[ \varphi_{o,i} = \sum_{j \in X} \varphi_{ij}, \quad \varphi_{d,j} = \sum_{i \in X} \varphi_{ij}. \]  

(5)

B. The Supply Side of an AMoD System

The supply side of an AMoD system describes the policy used by the fleet operator, i.e., the resources and routing decisions used to service demands. Formally, the supply is given by the tuple

\[ \text{Supply} : (N, f, \alpha). \]  

(6)

In (6), \( N \) is the number of vehicles in the fleet. Vehicles are homogenous and each may carry at most one passenger at a time. The function \( f : X \times X \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0} \) describes passenger-carrying trips, with \( f_{ij}(t) \) the flow of vehicles leaving \( i \), at time \( t \), carrying a passenger to \( j \). Similarly, \( \alpha : X \times X \times \mathbb{R} \rightarrow \mathbb{R}_{\geq 0} \) describes rebalancing trips, with \( \alpha_{ij}(t) \) the number of empty vehicles dispatched from \( i \), at time \( t \), to \( j \).

C. Performance Metrics of an AMoD System

How passengers are transported in an AMoD system is determined by the interaction between supply and demand. Accordingly, we define an AMoD system by the tuple \( (X, \lambda, \varphi, w, N, f, \alpha) \). In general, not every demand will be served immediately and not every vehicle will be in use at a given time, giving rise to queuing behavior at stations. Let \( Q_{ij}(t) \geq 0 \) denote the number of passengers waiting at station \( i \), at time \( t \), to travel to \( j \). Similarly, let \( V_i(t) \geq 0 \) denote the number of empty vehicles at station \( i \) at time \( t \). An important quality of any AMoD system is that the supply is able to keep up with the demand.

Definition 4.1: An AMoD system is stable if the number of demands in queue is bounded at all times, i.e.,

\[ \limsup_{t \rightarrow \infty} \sum_{i \in X} Q_{ij}(t) < \infty. \]  

(7)

We remark that if each \( Q_{ij} \) is served using a first-in first-out (FIFO) rule, which is typical for problem of this sort, then stability ensures each demand is served.

Subject to stability, we wish to route vehicles with an awareness for capital costs, operating costs, and the passenger experience. Capital costs are proportional to the total number of vehicles to be purchased, which we summarize by \( N \).

The operating costs is proportional to the total driving time. Assuming stability, each demand is transported from its origin to its destination and the only component of operating cost under the control of the fleet manager is the effort spent rebalancing vehicles. Accordingly, we define the rebalancing cost as

\[ C_r(\alpha) = \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{i,j \in X} \int_{0}^{T} \int_{t}^{\infty} \alpha_{ij}(t) \times \]  

\[ w_{ij}(t, \tau) \gamma_{ij}(t, \tau) d\tau dt. \]  

(8)
where $\gamma_{ij}(t, \tau)$ is the cost of an empty vehicle departing station $i$ at time $t$ and arriving at station $j$ at time $\tau$. Note that $C_q(\alpha)$ is linear in $\alpha$.

The passenger experience is related to the amount of time the average passenger spends waiting to enter a vehicle. Assuming any passenger’s time is as valuable as that of any other, we define the queueing cost as

$$C_q(f) = \limsup_{T \to -\infty} \frac{1}{T} \sum_{t, j \in X} \int_0^T Q_{ij}(t)dt,$$

where (9) $Q_{ij}(t) = Q_{ij}(0) + \int_0^t (\lambda(\tau)\varphi_{ij}(\tau) - f_{ij}(\tau))d\tau.$

Note that $C_q(f)$ is affine in $f_{ij}(t)$. The remainder of the paper is devoted to quantifying the tradeoffs, between $N$, $C_r$, and $C_q$, facing the fleet operator.

V. A FLUID MODEL OF AMoD SYSTEMS WITH PERIODIC TIME-VARYING DEMAND

This section presents a deterministic fluid model of passenger and vehicle movement in an AMoD system. The decision variables in the formulation are the supply-side quantities, i.e., $N$, $f_{ij}(t)$ and $\alpha_{ij}(t)$, which collectively constitute a routing policy. As mentioned, the travel demand $(\lambda(t), \varphi_{ij}(t))$ is periodic, with period $T_P$. We will consider periodic solutions, which correspond to equilibrium states for the system, when sampled at multiples of $T_P$.

Problem 5.1: Given an AMoD system with demand $(X, \lambda, \varphi, w)$ and rebalancing cost $\gamma$, find the routing policy $(N, f, \alpha)$ that solves the following linear program (LP) over the open simplex $\sum_{k=1}^3 a_k = 1$, $a_k > 0$.

$$\min_{N, f, \alpha} a_1 N + a_2 C_r(\alpha) + a_3 C_q(f),$$

s.t. $$Q_{ij}(t) = Q_{ij}(0) + \int_0^t (\lambda_{ij}(\tau) - f_{ij}(\tau))d\tau$$

$$V_i(t) = V_i(0) - \sum_{j \in X} \int_0^t (f_{ij}(\tau) + \alpha_{ij}(\tau))d\tau + \sum_{j \in X} \int_0^t (f_{ji}(\tau) + \alpha_{ji}(\tau))d\tau,$$

$$N = \sum_{i \in X} V_i(0) + \sum_{i, j \in X} \int_0^\infty \frac{1}{0} (f_{ji}(\sigma) + \alpha_{ji}(\sigma)) \times \frac{w_{ji}(\sigma, \tau)}{d\sigma}d\tau,$$

$$Q_{ij}(0) = Q_{ij}(T_P),$$

$$V_i(0) = V_i(T_P),$$

$N$, $f_{ij}(t)$, $\alpha_{ij}(t)$, $Q_{ij}(t)$, $V_i(t) \geq 0$, where the constraints on $Q_{ij}(t)$ holds for $i, j, i \neq j \in X$, $t \in [0, T_P]$ and those on $V_i(t)$ hold for $i \in X$, $t \in [0, T_P]$.

A few remarks are in order. First, the objective function is a convex combination of $N$, $C_r(\alpha)$, and $C_q(f)$, with each triple $(a_1, a_2, a_3)$ associated with a routing policy $(N, f, \alpha)$. The first two constraints specify the evolution of the customer and vehicles queues at stations, respectively. The constraints $Q_{ij}(t) \geq 0$ and $V_i(t) \geq 0$ ensure that passengers and vehicles that are not present at a station are not dispatched from that station. The number of vehicles in the system is expressed, at time $t = 0$, as the sum of the number of vehicles parked at stations plus the number that are traveling between stations. Finally, the periodic constraints on $Q_{ij}$ and $V_i$ ensure the system is stable according to (7), provided $Q_{ij}(0) < \infty$.

VI. THE CASE OF STATIONARY DEMAND

To reveal the design tradeoffs that materialize only when the demand is time-varying, this section considers the LP in Problem 5.1 for a stationary demand. For $\lambda = 0$ and $\varphi_{ij} = 0$, it follows that $Q_{ij}(0) = 0$, and $V_i(t) = 0$. In other words, there is no queueing and each passenger is picked up upon arrival. Dropping the time dependency, the LP reduces to

$$\min_{N, f, \alpha} a_1 N + a_2 C_r(\alpha) + a_3 C_q(f),$$

s.t. $$\lambda_{ij} = f_{ij}, \quad i, j, i \neq j \in X$$

$$\sum_{j \in X} (f_{ij} + \alpha_{ij}) = \sum_{j \in X} (f_{ji} + \alpha_{ji}), \quad i \in X$$

$$N = \sum_{i \in X} V_i + \sum_{i, j \in X} (f_{ji} + \alpha_{ji}) \overline{w}_{ji},$$

$$N, f_{ij}, \alpha_{ij}, Q_{ij}, V_i(t) \geq 0, \quad i, j, i \neq j \in X,$$

where $\overline{w}_{ij} = \int_0^\infty tw_{ij}(0, t)dt$ (11) is the average time required to travel from $i$ to $j$. Upon reducing $N, C_r$, and $C_q$ to the static case, the preceding LP turns out to be equivalent to a rebalancing policy originally reported in [13]. The following result begins the process of establishing this connection.

Proposition 1: Consider an AMoD system with static demand and $\gamma_{ij}(\sigma, \tau) = \sigma - \tau$ for $\sigma \geq \tau$ and zero otherwise. The objectives $N$ and $C_r(\alpha)$ are aligned. In other words, minimizing $N$ is equivalent to minimizing $C_r$ and vice versa.

Proof: The steady-state condition implies $\lambda_{ij} = f_{ij}$, ensuring $Q_{ij} = 0$, $C_q = 0$, and, in turn, the system is stable, i.e., (7) holds. Also, for the $\gamma_{ij}$ mentioned, $C_r(\alpha)$ in (8) reduces to

$$C_r(\alpha) = \limsup_{T \to -\infty} \frac{1}{T} \sum_{i, j \in X} \int_0^T \alpha_{ij} w_{ij}(0, \tau)d\tau d\tau$$

$$= \sum_{i, j \in X} \overline{w}_{ij} \overline{w}_{ij}. (12)$$

where $\overline{w}_{ij}$ is the average time required to travel from $i$ to $j$. Upon reducing $N, C_r$, and $C_q$ to the static case, the preceding LP turns out to be equivalent to a rebalancing policy originally reported in [13]. The following result begins the process of establishing this connection.

Proposition 1: Consider an AMoD system with static demand and $\gamma_{ij}(\sigma, \tau) = \sigma - \tau$ for $\sigma \geq \tau$ and zero otherwise. The objectives $N$ and $C_r(\alpha)$ are aligned. In other words, minimizing $N$ is equivalent to minimizing $C_r$ and vice versa.

Proof: The steady-state condition implies $\lambda_{ij} = f_{ij}$, ensuring $Q_{ij} = 0$, $C_q = 0$, and, in turn, the system is stable, i.e., (7) holds. Also, for the $\gamma_{ij}$ mentioned, $C_r(\alpha)$ in (8) reduces to

$$C_r(\alpha) = \limsup_{T \to -\infty} \frac{1}{T} \sum_{i, j \in X} \int_0^T \alpha_{ij} w_{ij}(0, \tau)d\tau d\tau$$

$$= \sum_{i, j \in X} \overline{w}_{ij} \overline{w}_{ij}. (13)$$

where $\overline{w}_{ij}$ is the average time required to travel from $i$ to $j$. Upon reducing $N, C_r$, and $C_q$ to the static case, the preceding LP turns out to be equivalent to a rebalancing policy originally reported in [13]. The following result begins the process of establishing this connection.

Proposition 1: Consider an AMoD system with static demand and $\gamma_{ij}(\sigma, \tau) = \sigma - \tau$ for $\sigma \geq \tau$ and zero otherwise. The objectives $N$ and $C_r(\alpha)$ are aligned. In other words, minimizing $N$ is equivalent to minimizing $C_r$ and vice versa.

Proof: The steady-state condition implies $\lambda_{ij} = f_{ij}$, ensuring $Q_{ij} = 0$, $C_q = 0$, and, in turn, the system is stable, i.e., (7) holds. Also, for the $\gamma_{ij}$ mentioned, $C_r(\alpha)$ in (8) reduces to

$$C_r(\alpha) = \limsup_{T \to -\infty} \frac{1}{T} \sum_{i, j \in X} \int_0^T \alpha_{ij} w_{ij}(0, \tau)d\tau d\tau$$

$$= \sum_{i, j \in X} \overline{w}_{ij} \overline{w}_{ij}. (13)$$
Additionally, the fleet size must satisfy
\[ N \geq \sum_{ij} (f_{ij} + \alpha_{ij}) w_{ij} = \sum_{ij} f_{ij} w_{ij} + C_r(\alpha). \]  

(14)
The term \( \sum_{ij} f_{ij} w_{ij} \) in (14) represents the number of vehicles transporting passengers in the network. Because \( f_{ij} = \lambda_{ij} \), the fleet operator has no control over this term and may influence \( N \) only through \( C_r \). Consequently, minimizing \( N \) is equivalent to minimizing \( C_r(\alpha) \).

Using Proposition 1, the LP in Problem 5.1 may be written as
\[
\min_{\alpha} \sum_{ij} \alpha_{ij} w_{ij} \\
\text{s.t.} \sum_{j \in X} (\lambda_{ij} + \alpha_{ij}) = \sum_{j \in X} (\lambda_{ji} + \alpha_{ji}), \forall i \in X.
\]

Note that, subject to a multiplicative constant, i.e., velocity, this LP is an EMD calculation. Here, the vehicle conservation constraint ensures the dropoff distribution, i.e., the supply of empty vehicles, is transformed into the pickup distribution, i.e., the demand for transport. The cost function ensures the ensuing shuffling of vehicles is done as efficiently as possible.

VII. Fleet Sizing to Avoid Queueing in Two-Station Networks

In general, AMoD systems are queueing systems where \( N \) is the number of servers, \( C_q \) is the service level, and \( C_r \) represents how aggressively servers are allocated to jobs. In queueing theory, analytic results typically assume something about the arrival or service processes, e.g., the arrival process is Markovian. Establishing connections between \( N, C_q \), and \( C_r \) is also likely to require refined knowledge of some subset of \( \lambda, \varphi_{ij}, w_{ij}, \gamma_{ij} \), and \( X \). However, the results in this section, which focus on the number of vehicles required to avoid passenger queueing, hold for general \( \lambda(t) \). The following result applies to a two-station network with deterministic travel times and illustrates one branch of analysis that can be conducted for AMoD systems. The subsequent result conveys that worst-case behavior with respect to this metric occurs for disjoint demands, suggesting demands with this structure may play an important role in defining the limits of achievable performance.

**Proposition 2:** Consider a two-station network with \( X = \{1, 2\} \) and \( \varphi_{12}(t) = 1, \varphi_{21}(t) = 0, t \geq 0 \). Assume \( w_{ij}(t, \tau) = \delta(\tau - t - t_{ij}) \) for \( t_{ij} > 0 \), where \( \delta \) is the Dirac delta function. Let \( t = t_{12} + t_{21} \) be the time required for a vehicle to travel from station 1 to 2 and back again. Let \( \hat{N} \) be the minimum number of vehicles required for \( C_q = 0 \). For any \( \lambda(t) \),
\[
\hat{N} = \max_{t \in \mathbb{R}} \int_{t-t}^{t} \lambda(t) dt.
\]

(15)

**Proof:** To begin, note that because \( \varphi_{12}(t) = 1, t \geq 0 \), there is no reason to hold empty vehicles at station 2. That is, there exists an optimal policy that sends every vehicle that drops a passenger off at station 2 immediately back to station 1. We assume such a policy for the remainder of the proof. Second, at some time, all \( \hat{N} \) vehicles will be traveling. To see why, note that if this were untrue, then the previous point implies that for all \( t \), station 1 maintains a stockpile of \( \hat{N} \geq 1 \) vehicles. However, reducing the fleet by \( \hat{N} \) vehicles maintains \( C_q = 0 \), which contradicts the optimality of \( \hat{N} \). Finally, when vehicles are not allowed to accumulate at station 2, and passengers are not allowed to accumulate at station 1, the total number of vehicle in operation at time \( t \) is given by
\[
\int_{t-t_{12}}^{t} \lambda(t) dt + \int_{t-t_{12}}^{t} \lambda(t) dt,
\]

(16)

where the first term represents the number of empty vehicle en route from station 2 to station 1, and the second term represents the number of passenger-filled vehicles en route from station 1 to station 2. The expression in (16) is equivalent to (15). Therefore, for bounded \( \lambda \), maximizing (16) over \( t \) is well-defined and yields \( \hat{N} \).

Intuitively, for a given \( \lambda(t) \), an AMoD system is under the greatest “stress” when supply and demand are disjoint, i.e., the set of origin and destination stations are mutually exclusive. In such a case, the set of supply and demand stations are a generalization of the system in Proposition 2, suggesting a disjoint two-station network may be indicative of worst-case behavior. In other words, for any \( (\omega_1, \omega_2, \omega_3) \), the worst-case cost for a given \( \lambda(t) \) is realized for \( X = \{1, 2\} \) and disjoint \( \varphi_{ij} \). The following results show this is true in the case of the minimum \( \hat{N} \) required for \( C_q = 0 \).

**Proposition 3:** For an AMoD system with stations \( X \), let \( X = P(X_1, X_2) \) denote that \( X \) can be partitioned into the non-empty subsets \( X_1 \) and \( X_2 \). Let \( \Phi_P \) denote the set of disjoint spatial demand profiles, i.e.,
\[
\Phi_P = \{ \varphi | \exists X_1, X_2 \text{ with } X = P(X_1, X_2) | \]
\[
\varphi_{ij} > 0 \Rightarrow i \in X_1, j \in X_2 \}.
\]

Let \( \hat{N}(\lambda(t), \varphi(t)) \) denote the minimum number of vehicles such that there exists a routing policy \( \{N, f_{ij}, \alpha_{ij}\} \) with \( C_q = 0 \). For any \( \lambda(t) > 0 \),
\[
\arg \max \{ \hat{N}(\lambda(t), \varphi_{ij}(t)) \} \in \Phi_P.
\]

(17)

\[\square\]

In words, (15) says that for a given \( \lambda(t) \), the minimum number of vehicles to avoid queueing is the largest, across the space of \( \varphi \), when the demand is disjoint.

**Proof:** To begin, consider \( \varphi_{ij} \notin \Phi_P \). Then, there must exist stations \( i, j, k \), as well as time \( t \) such that \( \varphi_{ij}(t) > 0 \) and \( \varphi_{jk}(t) > 0 \), i.e., \( j \) is a source and destination node. Because \( C_q = 0, f_{ij}(t) > 0 \) and \( f_{jk}(t) > 0 \),
\[\square\]
In this framework, snapping origin and destination points to the nearest station. The demand model is obtained by using a k-means clustering algorithm to obtain a network into fifteen minute intervals and aggregated demands for an initial user base for an AMoD system. We can imagine these individuals being representative of an twenty-four hour period by respondents. Collectively, we can form a national travel survey conducted in Singapore during 2012. The survey lists all trips, i.e., origin location, destination location, and departure time, taken over a twenty-four hour period by respondents. Collectively, we can imagine these individuals being representative of an initial user base for an AMoD system.

To solve Problem 5.1 numerically, we discretized time into fifteen minute intervals and aggregated demands using a k-means clustering algorithm to obtain a network of thirty stations. The demand model is obtained by aggregating trips into fifteen minute windows and mapping origin and destination points to the nearest station. In this framework, \( f_{ij}(k) \) and \( \alpha_{ij}(k) \) represent the total passenger-filled and empty vehicles to be dispatched from station \( i \) to \( j \) during the interval \( k \in [0, 24 \times 4] \). The simulation assumes vehicles travel at a constant speed of 30km/h and move directly between stations. Upon snapping origin and destination points to the nearest station, and discarding trips that start and end the same node, we were left with 39,273 trips.

The arrival of passengers and movement of vehicles was simulated using a step size of fifteen seconds. Because this is much smaller than the timescale used to solve for optimal flows, \( f_{ij}(k) \) and \( \alpha_{ij}(k) \) serve as upper bounds on the number of passenger-filled and empty vehicles that can be dispatched from \( i \) to \( j \) in interval \( k \).

Because travel and inter-arrival times are not multiples of fifteen minutes, the resulting simulation can appear bursty because passenger and vehicle trips are held at stations until the next interval begins. We have found this effect, which is an artifact of discretizing a continuous problem and then reverting to a finer timescale for simulation, can be resolved by introducing the cumulative flow budgets \( F_{ij} \) and \( A_{ij} \). Here, \( F_{ij} \) is the total number of passenger-filled vehicles that can be sent from \( i \) to \( j \) by the end of the \( k \)-th interval, and is given by

\[
F_{ij}(k) = \begin{cases} 
  f_{ij}(0) + f_{ij}(1) & \text{if } k = 0 \\
  F_{ij}(k-1) + f_{ij}((k-1) \mod T_P) & \text{if } k > 0.
\end{cases}
\]

\( A_{ij} \) is defined in a similar way for empty vehicles. These revised quantities are used in the simulations that follow.

To illustrate the realizable Pareto optimal \((N,C_r,C_q)\), we considered a range of \( N \), and, for each \( N \), solved Problem 5.1 for a range of \((a_2,a_3)\), \( a_2 + a_3 = 1 \). Figure 1, reveals the operating points achievable for the associated network and demand in Singapore.

The fleet operator may use curves of this form to determine an appropriate routing policy. For example, given a vehicle budget and bound on the permissible level of queuing, the operator can use Figure 1 to determine which operating points satisfy the listed criteria. Alternatively, if none of the policies are acceptable, an AMoD system may not be appropriate for the market. In either case, solutions to Problem 5.1 inform the fleet operator of the performance that can be achieved.

**VIII. Simulation Results**

In this section, we show how a fleet operator can, by solving Problem 5.1, select a routing policy tailored to their needs. To ground our study, we consider a hypothetical AMoD installation in Singapore, and base the demand on recorded trip data. We begin by briefly discussing our data source and outlining our methodology, before analyzing the simulation results.

The demand model was populated using data from a national travel survey conducted in Singapore during 2012. The survey lists all trips, i.e., origin location, destination location, and departure time, taken over a twenty-four hour period by respondents. Collectively, we can imagine these individuals being representative of an initial user base for an AMoD system.

This paper provided a fluid-based optimization framework that expresses tradeoffs between fleet size, rebalancing effort, and queuing effects in terms of passenger and vehicle flows for a shared mobility system with time-varying demand. This model recaptures a previously reported rebalancing policy in the special case of static
demand. Additionally, we quantified the number of vehicles necessary to avoid passenger queueing, beginning with a two-station network, before showing that, for a general network, this quantity is greatest when the demand is disjoint. Finally, simulations, using trip data from Singapore, demonstrated how our model could be used by a fleet operator to select a vehicle routing policy.

Looking forward, there are a number of natural outlets for future work. Theoretically, we have only scratched the surface of quantifying fundamental relationships between $N$, $C_r$, and $C_q$. To delve deeper, it may prove useful to focus on $\lambda(t)$ and $\varphi_{ij}(t)$ that possess special properties. For example, concentrate on demands with the property that for each trip from $i$ to $j$, there is a corresponding trip from $j$ to $i$ later in the day. Demands with this property are not only representative of observed mobility patterns, but their added structure may spark more compelling tradeoffs among the metrics of interest.

Additionally, our formulation ignores the congestive effects of traffic. In reality, the decision to route vehicles on a given link should take into account the total number of vehicles scheduled to use the link. Moreover, congestion may prompt temporal shifts in rebalancing trips to mitigate traffic along popular travel corridors during peak hours. The implications of these effects on optimal routing strategies are potentially profound and the inclusion of congestion would add considerable credence to our optimization framework.

References


