IFAC Journal of Systems and Control 22 (2022) 100210

Contents lists available at ScienceDirect

IFAC Journal of Systems and Control

journal homepage: www.elsevier.com/locate/ifacsc

Fleet sizing and charger allocation in electric vehicle sharing systems

Yuntian Deng [a,*], Abhishek Gupta [a], Ness B. Shroff [a,b]

[a] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA
[b] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 10 April 2020
Received in revised form 1 September 2022
Accepted 23 September 2022
Available online 1 October 2022

Keywords:
Electric vehicles
Sharing economy in transportation
Closed queueing networks

ABSTRACT

In this paper, we propose a closed queueing network model for performance analysis of electric vehicle sharing systems with a certain number of chargers in each neighborhood. Depending on the demand distribution, we devise algorithms to compute the optimal fleet size and number of chargers required to maximize profit while maintaining a certain quality of service. We show that the profit is concave with respect to the fleet size and the number of chargers at each charging point. If more chargers are installed within the city, we show that it can not only reduce the fleet size, but it also improves the availability of vehicles at all the points within a city. We further show through simulation that two slow chargers may outperform one fast charger when the variance of charging time becomes relatively large in comparison to the mean charging time.

1. Introduction

Many governments across the world are emphasizing the decarbonization of transportation to curb greenhouse gas emissions and pollution associated with transportation industry. With the rise in sharing economy within transportation sector, there is a shift from personally-owned modes of transportation to shared vehicles (using ride-hailing services, bikes, and scooters, etc.). These service providers are now commonly referred to as transportation network companies (TNCs). TNCs provide on-demand transportation services to the passengers, increases vehicle utilization, and enhances overall convenience to the passengers. Incorporating electric vehicle within TNCs is widely considered to be a solution to achieve long-term sustainable transportation objectives.

Given the high annual miles traveled by vehicles in shared fleets, Pavlenko, Slowik, and Lutsey (2019) estimates that the per mile operating cost of all battery electric or hybrid vehicle fleets is much lower than that of the conventional ones. Even without the current purchasing incentives, long range battery electric vehicles (BEVs) will become the most economically attractive technology for ride-hailing operations by 2023–2025 (assuming the cost of battery packs go down by 35%). Another recent research suggests that the operating and ownership expenses of electric vehicles with high annual miles driven are significantly lower than those of conventional vehicles (Weldon, Morrissey, & O'Mahony, 2018). However, charging electric vehicles takes a non-trivial amount of time (depending on the state of charge of the vehicle). Moreover, the cost of installing and maintaining charging infrastructure is substantial (installation of a charging station could cost anywhere between $ 10–50 thousand and reserving parking spots for electric vehicles could be costly in high population density areas). Consequently, for the adoption of electric vehicle sharing system, it is important to determine based on the demand distribution:

(1) What should be the optimal fleet size, since a large fleet of EVs results in improved availability, reliability, and better quality of service, but it also costs more to maintain.
(2) What should be an optimal number of chargers at stations across the city. A large number of charging infrastructure would improve availability and quality of service, but it has high recurring costs.

We formulate these two problems as optimization problems in which the movement of the vehicles across a city is modeled using a closed queueing network model. We devise algorithms to solve these optimization problems. We also shed some light on the nature of charger (fast vs. slow) needed for improving the quality of service.

1.1. Literature review

Vehicle sharing systems offer customers to rent car for a short period of time (Ataç, Obrenović, & Bierlaire, 2021). These vehicle sharing systems vary in terms of demand type (one-way Kaspi, Raviv, & Tzur, 2014 vs. round-trip Lee, Quadrifoglio, Meloni, et al., 2016), parking location (free-floating Weikl & Bogenberger, 2015 vs. station-based Lee et al., 2016), reservation horizon (advanced Kaspi et al., 2014 vs. last-minute Weikl &

* Corresponding author.
E-mail addresses: deng.556@osu.edu (Y. Deng), gupta.706@osu.edu (A. Gupta), shroff.11@osu.edu (N.B. Shroff).

https://doi.org/10.1016/j.ifacsc.2022.100210
2468-6018/

Bogenberger, 2015), relocation capability (Illgen & Höck, 2019) (passive vs. active vs. no relocation), and vehicle engine type (internal combustion engine Weikl & Bogenberger, 2015 vs. electric Boyacı, Zografos, & Geroliminis, 2015). In this paper, we focus on strategic decisions in the one-way station-based last-minute electric vehicle sharing system.

A significant amount of work has focused on the strategic decisions for electric vehicle service. We refer the readers to a survey (Shen, Feng, Mao, & Ran, 2019) and review (Kumar & Alok, 2020) for a comprehensive overview. One important decision is strategically planning EV charging infrastructures, to meet local charging demands and reduce social costs. Most works focus on the flow-based models (Hodgson, 1990), where demand is proportional to the traffic flow that will pass by the charging facilities. For example, Huang, Li, and Qian (2015) developed a multipath charging infrastructure location model, where travelers could deviate from the shortest path in an origin–destination pair and EVs with limited range can be charged en route to their destinations.

Compared with consumer passenger vehicles, charging infrastructures for shared taxi fleet is more challenging, due to the stochastic nature of the problem. e.g., the arrival of demand (EV) is stochastic and the charging time varies from vehicle to vehicle. Jung, Chow, Jayakrishnan, and Park (2014) proposed a new charger location algorithm for shared taxi fleet, which features stochastic demand occurring dynamically over time and charging queueing delay for the service fleet. Yang, Dong, and Hu (2017) presented a data-driven optimization approach to allocate chargers for BEV taxi. An $M/M/x/s$ queueing model is used to capture the charging congestion and demand arrivals. However, these works are based on open queueing network (EVs arrive at the queueing network, receive charging, and departure), where the number of vehicles is ignored and other operations for shared EV are not captured, such as waiting for passengers and traveling on roads. Such Markov chain based models are also utilized in car-sharing (Repoux, Kaspi, Boyacı, & Geroliminis, 2019) and bike-sharing systems (Raviv & Kolka, 2013).

Closed queueing network (Baskett, Chandy, Muntz, & Palacios, 1975), overcomes above issues through fixing the number of EV in the system and leveraging the interactivity in the system, i.e. EVs are modeled as picking up passengers, traveling, (possible) charging, and then picking up new customers. George and Xia (2011) first proposed a closed queuing network based model for a vehicle rental system, but not for EVs. It considers optimal fleet sizing with additional constraints on quality of service parameters such as availability. Fanti, Mangini, Pedroncelli, and Ukovich (2014) considered three different types of electric vehicles in a closed queueing network model for EV sharing system: fully charged, partially charged and out of charge. Iglesias, Rossi, Zhang, and Pavone (2019) cast an autonomous mobility-on-demand system within the framework of BCMP closed queueing networks, which can capture both congestion effects and vehicle charging. However, they assume that the charging infrastructure at each site is unlimited, which is not realistic. Unlike the simulation method, we also provide theoretical results on the properties of the closed queueing EV sharing model, which could provide high-level insights for strategic decisions.

1.2. Key contributions of this paper

Our research is motivated by the real world constraint of having finite number of chargers in the city. First, we model the one-way station-based EV sharing systems using a closed BCMP queueing network model with finite number of chargers. Through this model, we are able to highlight the challenges in strategic decisions, such as fleet sizing and allocating finite number of chargers in the city to maximize profit, as well as charger selection. Secondly, since different neighborhoods may have different requirements on vehicle availability and may have different numbers of chargers, we propose optimization formulations to determine the optimal fleet size and charger allocation to maximize profit. Concavity of profit with respect to fleet size and the number of chargers is established when the charging time distribution is exponential. We further devise an algorithm to solve the charger allocation problem in an iterative fashion. Thirdly, we show that under a fixed budget, two slow chargers outperform one fast charger, whose charging time is halved compared with slow chargers, if the variance of charging time becomes relatively large. An approximation method is proposed for general passenger inter-arrival time. Finally, large-scale simulations validate our theorems and approaches.

1.3. Outline of the paper

The rest of the paper is organized as follows. The system model is presented in Section 2. Section 3 reviews some results on invariant distribution and throughput of closed queueing networks. The optimal fleet sizing problem is formulated and solved in Section 4. In Section 5, we devise a greedy approach based heuristic algorithm for charger allocation based on marginal allocation algorithm of Fox (1966). The comparison between one fast and two slow chargers is discussed in Section 6. To reduce the computational complexity for large-scale simulations, the mean value analysis is reviewed in Section 7. We also discuss in this section an approximation approach for computing the stationary probability distribution of the system for the case of general passenger inter-arrival time distribution. Section 8 presents the numerical results for both optimization problems for a 60-node city network. Finally, Section 9 concludes the paper and presents directions for future research.

2. System modeling

In this section, we model an electric vehicle sharing system using the framework of Baskett, Chandy, Muntz, and Palacios (BCMP) closed queueing network (Baskett et al., 1975). This model for vehicle sharing system is adopted in Iglesias et al. (2019). Suppose there are $M \in \mathbb{N}$ electric vehicles in one city, which can offer services to passengers and are routed around to serve stochastic demands at different places. Similar to George and Xia (2011), from a virtual service view, we model this system as a closed network with respect to vehicles, i.e., vehicles are routed within this network and receive 'service' at different nodes. The number of vehicles remains constant and there is no vehicle entering or leaving the network.

As shown in Fig. 1, we use three different queues to model three processes: departure, charging and travel. To be specific, each station comprises a single-server queue (SS) node (departure) and a finite-server queue (FS) node (charging). The travel between two stations is considered as an infinite-server queue (IS) node. In the following, we explain the system from view of the passengers and the electric vehicles, respectively.

2.1. Passenger

Assume that there are several stations within a city. Each station has a departure point (pick-up node) and a charging point (drop-off node). We denote the set of departure points as S and the set of charging points as F. At each departure point $i \in S$, passengers arrive according to a Poisson arrival process with rate $\alpha_i > 0$. If there is at least one electric vehicle waiting at departure point i, the passenger will take the first electric vehicle in line and start traveling without any waiting time. If there is no

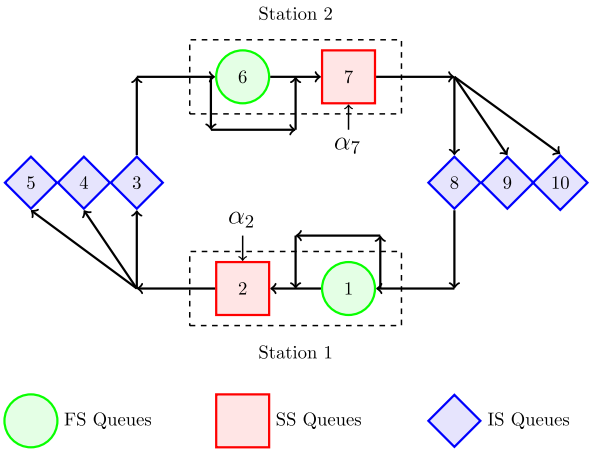


Fig. 1. Electric Vehicle Sharing System, in which finite server queues denote charging stations, single server queues denote the passenger pickup stations, and infinite server queues represent road networks.

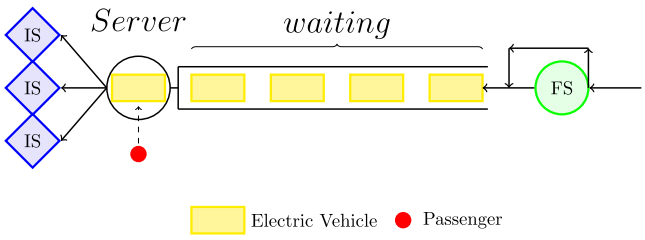


Fig. 2. Single Server (SS) queue: electric vehicles queue up, waiting for incoming passengers. Passenger will always pick the first vehicle in line if there are some vehicles waiting.

electric vehicle at node i, then the passenger leaves the system immediately and takes some other mode of transportation. Thus, there is a "passenger loss" in this situation; the passengers do not form a queue at node i.

Before departing from point $i \in S$, each passenger selects his/her destination as the charging point $j \in F$ with probability p_{ij}, where $\sum_{j\in F} p_{ij} = 1$ for each $i \in S$. We assume that the travel time from node i and to j follows a general distribution with mean T_{ij}.

Passengers exit this system as soon as they arrive at the charging point j. For example, if a passenger wants to travel from station 1 to station 2, as shown in Fig. 1, he/she will depart from node 2 and enter node 3. Once the passenger leaves node 3 and arrives at node 6, then the one-way mobility-on-demand service is finished. If the vehicle does not need charging (discussed in Section 2.2.3), it will be parked in lots. If the vehicle needs to be charged, then one staff at this station will handle the connection/disconnection procedure and there are two possible cases. Case I: there is at least one charger available and the staff will connect it with one empty charger. Case II: all chargers are occupied and the vehicle will be put in the waiting stage. Once a vehicle is fully charged, the staff will unplug the vehicle and plug in one of the waiting vehicles. We assume the charging cable is long enough and there is sufficient parking space.

2.2. Electric vehicles

Electric vehicles are routed among three types of queues in the network according to probability r_{ij} (defined later by (1)). We assume that the transfer from one queue to another is instant. From a virtual service view, electric vehicles form queues and receive services when they are waiting for passengers (SS), traveling between stations (IS), and charging at charging points (FS). Charging can also be skipped according to a certain probability, and in this case, electric vehicles directly go to the departure points after leaving the IS nodes.

2.2.1. Single Server (SS) queues — Departure

At each departure point, vehicles queue up to wait for the arrival of the next passenger. If there is no vehicle at the departure point, then any passenger who arrives would leave immediately. We view node $i \in S$ as a First-Come-First-Serve (FCFS) single server (SS) queue, whose service rate is the passenger's arrival rate at station i: α_i, i.e., the service time is exponentially distributed with mean $\frac{1}{\alpha_i}$, which is the same as the inter-arrival time of passengers at this point. As shown in Fig. 2, if there are at least two electric vehicles at the departure node $i \in S$, one vehicle is being served in the server of the queue and others are waiting in the queue.

Once there is an arrival of a new passenger, the vehicle in the server will finish its service and leave this node carrying one passenger, i.e., the arriving passenger will pick the first vehicle in the line. When the server is idle, the second vehicle will enter the server and start its service. This is similar to an airport taxi service when there is only one line. The first taxi is always on the server and it departs once a passenger arrives, the second taxi becomes the first one after the departure following the first-come first-serve discipline.

2.2.2. Infinite Server (IS) queues — Travel

After departing from node $i \in S$, each vehicle will go towards its destination $j \in F$ selected by passenger with probability p_{ij}, where $\sum_j p_{ij} = 1$ and $p_{ii} = 0$. We use an infinite-server (IS) queue connecting the origin and destination to model the travel time of passengers. We assume that the travel time is independent across all the passengers and follows a general distribution with mean T_{ij}, which is associated with the distance between departure point i and charging point j.

2.2.3. Direct path — No charging

After the passenger is dropped off at the destination, the electric vehicle may need to be charged. Accordingly, we assume that with probability $\bar{p}_{ij}$, the vehicle decides to be charged at node $j \in F$ for $i \in I$, and with probability $1 - \bar{p}_{ij}$, it goes directly to the following SS queue without waiting or charging.

2.2.4. Finite server (FS) queues — Charging

If the vehicle decides to be charged, we use an FCFS finite server (FS) queue to model the charging process at charging point $j \in F$. As both chargers and spaces are limited at each station, we assume that the maximum number of vehicles charged simultaneously is v_j at charging point j, i.e. there are v_j chargers at charging point j. All vehicles that decide to be charged at that node forms an FCFS queue to wait for charging. To simplify the analysis, we assume that the charging time follows an exponential distribution with mean t_j for $j \in F$. For a general charging time distribution, the exact solution is still an open problem (Gupta, Harchol-Balter, Dai, & Zwart, 2010) in queueing theory. We provide an insight for approximating it in Section 7.3. After the charging process is over, charged electric vehicles will enter the following single server queue to wait for the passengers.

2.3. Closed queueing network

In this network, we have considered three types of nodes: Single server queue (S), Infinite server queue (I) and Finite server queue (F). Let $\mathcal{N} = S \cup I \cup F$ denote the set of all nodes, and

denote $N = |\mathcal{N}|$. For each node $i \in \mathcal{N}$, let Parent(i) be the direct origin of node i, i.e., as shown in Fig. 1, Parent(7) is node 6 and Parent(2) is node 1.

The routing matrix of vehicles between nodes can be written as follows:

$$r_{ij} = \begin{cases} p_{il}, \ i \in S, j \in I, l \in F, \\ \qquad i = \text{Parent}(j), j = \text{Parent}(l) \\ \bar{p}_{ij}, \ i \in I, j \in F, i = \text{Parent}(j) \\ 1 - \bar{p}_{ik}, \ i \in I, k \in F, j \in S, \\ \qquad i = \text{Parent}(k), k = \text{Parent}(j) \\ 1, \ i \in F, j \in S, i = \text{Parent}(j) \\ 0, \ \text{otherwise} \end{cases} \tag{1}$$

where the first case means that, after selecting the destination $l \in F$ with probability p_{il}, passengers enter the associated roads and begin traveling at node $j \in I$ following departure from node $i \in S$. The second case indicates that vehicles will choose to charge at charging point $j \in F$ with probability $\bar{p}_{ij}$ for $i \in I$ after exiting the roads $i \in I$. The third case denotes that vehicles will move directly to the departure point $j \in S$ after exiting the roads $i \in I$ with probability $1 - \bar{p}_{ik}$, which skips charging at $k \in F$. The fourth case indicates that all vehicles will move to the departure points S if they finish the charging process within the same station.

When there are $n_i \in \{0, \ldots, M\}$ vehicles at node $i \in \mathcal{N}$, the service rate at each node (the average number of vehicles finishing service and leaving this node per unit time) is as follows:

$$u_i(n_i) = \begin{cases} \alpha_i & n_i \geq 1, \ i \in S \\ 0 & n_i = 0, \ i \in S \\ \frac{n_i}{T_{jl}} & j \in S, i \in I, l \in F, j = \text{Parent}(i), i = \text{Parent}(l) \\ \frac{\min\{n_i, v_i\}}{t_i} & i \in F \end{cases} \tag{2}$$

where the first and second cases mean that, if there is at least one vehicles at station, the arrival and departure process of passengers, with rate α_i, and does not depend on the number of vehicles, n_i, at this node. The third case indicates that all travel times are independent from each other and T_{jl} is the mean travel time from departure point $j \in S$ and charging point $l \in F$. The fourth case means that if the number of vehicles willing to be charged n_i is larger than the number of chargers v_i at charging point $i \in F$, they need to form a FCFS queue to wait until a vehicle finishes charging and the charger becomes available, while t_i is the average charging time at charging point $i \in F$.

3. Closed queueing network analysis

In this section, we introduce some results in a closed queueing network and in particular, the BCMP network. In our model from the last section, SS queues and FS queues fall into the type-I queues and IS queues belong to the type-III queue in BCMP network (Gelenbe & Pujolle, 1998). Therefore, this model falls into the class of closed BCMP network, which has the product-form solution to the stationary distribution, because these queues are quasi-reversible (Balsamo, 2000).

Given the fleet size M, i.e. there are M electric vehicles routing within the network and no electric vehicle enters or leaves the system, the associated continuous-time Markov process has the following space

$$\mathcal{S} = \left\{(n_1, n_2, \ldots, n_N) : \sum_{i=1}^{N} n_i = M, n_i \in \mathbb{N}\right\},$$

where n_i is the number of vehicles at node $i \in \mathcal{N}$. Since the transition from one node to another is instant in our model, every vehicle must be at one node $i \in \mathcal{N}$.

Let $\lambda = (\lambda_1, \ldots, \lambda_N)$ denote the relative throughput at node $i \in \mathcal{N}$, which is defined as the relative average number of vehicles passing through the node per unit time. Since there are a fixed number M of vehicles routing among the nodes, we have the following constraint (global balance equations):

$$\lambda_i \sum_{k \in \mathcal{N}} r_{ik} = \sum_{j \in \mathcal{N}} \lambda_j r_{ji}, \ \forall i \in \mathcal{N}, \tag{3}$$

where the probability of vehicle routing from node i to node j is r_{ij} in (1). With another constraint $\sum_{i \in \mathcal{N}} \lambda_i = 1$, we can find the unique solution to (3) with respect to λ, which is also called visit ratio (Balsamo & Marin, 2007, p.53).

It now follows from Baskett et al. (1975) that the stationary probability distribution of the resulting continuous time Markov process $P(n_1, n_2, \ldots, n_N)$ has the following product form :

$$P(n_1, \ldots, n_N) = \frac{1}{G(M)} \prod_{i=1}^{N} \frac{\lambda_i^{n_i}}{\prod_{k=1}^{n_i} u_i(k)}, \tag{4}$$

where $G(M)$ is the normalization constant in order to make its summation equal to one. A computational method to compute $G(M)$ is discussed in Section 7.1.

From operational perspective, throughput and availability are the key performance indicators for the overall system. Throughput captures how many passengers are served per unit time. Availability is defined as the probability that at least one vehicle is available at the departure point. Due to the product form of the invariant distribution, these two quantities can be computed in closed form, and the expressions are given in the following lemma.

Lemma 1. *In a closed BCMP queueing network with M vehicles and N nodes, the throughput and availability are as follows*

1. *The throughput of each node $i \in \mathcal{N}$ (the average number of vehicles passing through node i per unit time) is*

$$\Lambda_i(M) = \lambda_i \Lambda(M), \tag{5}$$

where the system throughput of the network is

$$\Lambda(M) = \frac{G(M-1)}{G(M)}. \tag{6}$$

2. *The availability at departure points S, i.e., the probability that node $i \in S$ has at least one vehicle is*

$$A_i(M) = P\{n_i \geq 1\} = 1 - P\{n_i = 0\} = \frac{\lambda_i}{\alpha_i} \Lambda(M). \tag{7}$$

Proof Sketch: The throughput in closed queue network is given by Serfozo (2012, p. 27) where λ_i is the visit ratio defined in (3) and $G(M)$ is the normalization factor defined in (4) and (17). The availability is defined as the probability of at least one vehicles at departure point $i \in S$, which equals to $1 - P\{n_i = 0\}$. The probability of zero vehicles at departure point $i \in S$ is computed in Lavenberg (1983, p. 128). We refer the reader to Appendix A for further details.

In (7), with a high vehicle arrival rate λ_i (high supply) and a low passenger arrival rate α_i (low demand) at departure point i, the service availability $A_i(M)$ will be relatively high. As defined before, we assume that there is a 'passenger loss' if there is no electric vehicle at departure point when a passenger arrives, i.e., the passenger leaves this system and try other modes of transportation. We show below that the probability of loss of a passenger at any time is related to the notion of availability introduced above.

Lemma 2. *If the fleet size is M, then the probability that there is a "passenger loss" at departure point $i \in S$ is $1 - A_i(M)$, where $A_i(M)$ is defined in Eq. (7).*

Proof. From Poisson arrival see time average (PASTA) (Wolff, 1982), the probability of the state as seen by an outside random observer is equal to the probability of the state seen by an arriving passenger under Poisson arrival. Recall that $1 - A_i(M)$ is the probability that an outside observer will find no vehicle at departure point $i \in S$. Due to PASTA property, this is also the probability that a newly arrived passenger will find no vehicle at departure point i. This is precisely the probability of a passenger loss at node i. □

In the following, we discuss the insensitivity property in the product form networks (Balsamo, 2000, Section 3.4), which shows that the stationary distribution of the network $P(n_1, n_2, \ldots, n_N)$ does not depend on the variance of service time distribution at infinite server nodes (IS), i.e., if the variance of travel time distribution between stations is changed, the average performance metrics (throughput, average waiting time, average queue length) remains the same.

Lemma 3. *Consider a closed BCMP network introduced above. Let $\xi = (\xi_i)_{i\in I}$ denote the travel time distribution of the vehicles in the infinite server node $i \in I$, and let $\bar{t}_{\xi_i}$ be the mean of the distribution ξ_i. Let $P_\xi(n_1, n_2, \ldots, n_N)$ denote the corresponding stationary distribution. If ξ' is another travel time distribution such that $\bar{t}_{\xi_i'} = \bar{t}_{\xi_i}$, then $P_\xi = P_{\xi'}$.*

Proof. The result follows from Chandy, Howard, and Towsley (1977, Corollary 4.1 & Theorem 6) via station balance. This property holds because units receive service immediately upon entering the queue and their wait times are zero. □

A consequence of the above result is that the stationary distribution is dependent only on the mean of the service time distribution for infinite server queues; the precise distribution does not matter. As the average performance metrics (throughput, average waiting time, average queue length) can be derived from stationary distribution P, they are also independent of service time distribution for infinite server queues.

Furthermore, we can extend the insensitivity property to finite server queues (FS) if the number of vehicles M is less than or equal to the number of chargers at finite server queues v_i. Intuitively, when the condition $M \le v_i$ holds, the queue i behaves as same as an infinite server queue, therefore the stationary state distribution only depends on the mean service time at finite server queue i.

Lemma 4. *In closed BCMP network, if $M \le v_i$, i.e., the number of vehicles in the network is equal to or less than the number of charger at node $i \in F$, then the stationary state distribution $P(n_1, n_2, \ldots, n_N)$ depends on the service time distribution at finite server queue i only through its mean.*

Proof. This lemma follows from Baskett et al. (1975, p.250 Condition 3) and its Section 4.1, where only the mean service times appear in $P(n_1, n_2, \ldots, n_N)$ and any service time distribution with the same mean yields the same results as exponential service time distribution. □

4. Optimal fleet sizing

From the view of a car-sharing operator, one critical variable is the size of an electric fleet, before launching service in one city. In this section, we want to develop a profit maximization problem with operating cost, by controlling the fleet size, while maintaining a certain quality of service. Ref. George and Xia (2011) considers such a fleet sizing problem for gasoline vehicles with exponential travel time distribution. We extend it to a more general case for EVs with any travel time distribution, a finite number of chargers, convex operating cost function and location-specific availability requirements.

As service providers can only make money when vehicles are traveling, we model its total revenue per unit time as $\sum_{i\in I} \Lambda_i(M) z_i$, where z_i is the revenue per-service (one-way charge) when vehicles are in IS nodes $i \in I$. $\Lambda_i(M)$ is the throughput of node i, i.e., the average number of service finished at node i per unit time. Besides, we define the operating cost (salary, maintenance, etc.) per unit time as a convex (including linear) increasing function $g(M)$ with respect to the fleet size M.

As there are various requirements of availability at different places (e.g., high availability at airports and downtown), we define $\epsilon = (\epsilon_1, \ldots, \epsilon_s)$ as the quality of service requirement in the system. At each departure point $i \in S$, the availability $A_i(M)$ defined in (7) is greater than or equal to $1 - \epsilon_i$.

From a steady-state view of the system, we want to maximize the profit by controlling the fleet size M, while maintaining a certain quality of service $A_i(M)$. The optimization problem can be formulated as follows:

$$\max_{M\in\mathbb{N}} f(M) = \sum_{i\in I} z_i \Lambda_i(M) - g(M) \tag{8}$$

$$s.t.\ A_i(M) \ge 1 - \epsilon_i,\ \forall i \in S \tag{9}$$

In the next two lemmas, we show that the objective function f is concave and that the above optimization problem is feasible. For a function $g : \mathbb{N} \to \Re$ defined over the space of natural numbers, it is said to be concave (Shanthikumar & Yao, 1988b) if

$$f(M) + f(M+2) \le 2f(M+1) \text{ for all } M \in \mathbb{N}.$$

Lemma 5. *The objective function $f : \mathbb{N} \to \mathbb{R}$ is concave in M.*

Proof Sketch: For the system under exponential travel time distribution without the charging stations, this result is established in George and Xia (2011, Theorem 2, p. 202). We show that essentially the same argument holds for our case with the charging stations and any travel time distribution in Appendix B, where we add the FS queue and extend the travel time distribution from exponential to any distribution.

In the following lemma, we show that, if there are more vehicles in the system, the availability at every departure point will increase.

Lemma 6. *The availability function $A_i(M)$ at each departure point $i \in S$ is non-decreasing with M.*

Proof. From the first part of proof in Lemma 5 and (7) $A_i(M) = \frac{\lambda_i}{\alpha_i}\Lambda(M)$, we find that $A_i(M)$ is non-decreasing with M. □

Therefore, if there exists M_{ϵ_i} such that $A_i(M) \ge 1 - \epsilon_i$ holds for $M \ge M_{\epsilon_i}$, let $M_\epsilon = \max_{i\in S} M_{\epsilon_i}$, we can conclude that the constraint (9) is satisfied for all $M \ge M_\epsilon$.

Theorem 7. *If $g(M) \to \infty$ as $M \to \infty$, the optimization problem above either has a unique solution or has multiple adjacent solutions.*

Proof. Take the backward discrete derivative as

$$\Delta f(M) = f(M) - f(M-1)$$

As $\Lambda(M)$ is upper bounded from (7) and $g(M) \to \infty$ as $M \to \infty$, we have $f(+\infty) \to -\infty$ since $f(M) = \Lambda(M)\sum_{i\in I} z_i\lambda_i - g(M)$.

Since f is concave from Lemma 5, $f(M)$ is decreasing when M is sufficiently large, then there exists at least a critical point, such that either (i) $\Delta f(M_1) > 0$ and $\Delta f(M_1+1) < 0$, or (ii) $\Delta f(M_2) = f(M_2+1) = \cdots = f(M_2+k-1) = 0$ for a constant k. Then $M^* = M_1$ in the first case, or $M^* = [M_2-1, \ldots, M_2+k-1]$ in the second case. □

From the solution provided in the proof above, we find that the optimal fleet size M^* is determined by various parameters: routing probability matrix r_{ij}, service rate $u_i(n_i)$, revenue per service z_i, fleet operating cost $g(M)$ and quality of service requirement ϵ. In Section 8, we show through numerical simulation the effect of these parameters on the optimal fleet size. We summarize the procedure to find the optimal fleet size in Algorithm 1.

Algorithm 1: Optimal Fleet Sizing

Input : $f(M)$, $A_i(M)$, and ϵ_i for all $i \in S$, left offset $n = 0$, right offset $k = 1$
Output: M^*

```
for M ← 1 to ∞ do
    if A_i(M) ≥ ε_i, ∀i ∈ S then
        if Δf(M) = 0 then
            if A_i(M − 1) ≥ ε_i, ∀i ∈ S then
                n = 1
            end
            for k ← 1 to ∞ do
                if Δf(M + k) < 0 then
                    return [M − n, M + k − 1]
                end
            end
        end
        if Δf(M + 1) < 0 then
            return M;
        end
    end
end
```

Remark 8. This algorithm provides a line search rather than bisection to utilize the efficient mean value analysis in Section 7.2 for availability and avoid large computation for $f(M)$ when availability condition is not satisfied.

5. Charger allocation

In practice, there are usually very limited spaces for charging in the downtown area (the rent is high) and building charging infrastructure takes a non-trivial amount of money and time. As a result, the service provider needs to decide on the location of the charging stations and the number of chargers to be installed at each charging station. Intuitively, if more chargers are built, electric vehicles will spend less time waiting or driving around looking for unoccupied chargers, which leads to more availability. On the other hand, building and operating more chargers will increase operating costs. Therefore, there is a trade-off between quality of service and operating cost.

We model it as a profit maximization problem, by controlling V, where $V = (v_1, v_2, \ldots, v_f)^T$ is the vector of the number of chargers at each charging point $i \in F$. Throughout this section, we fix the fleet size to M and consider the throughput and availability as a function of V, i.e. $\Lambda(V)$ and $A(V)$ are short for $\Lambda(V, M)$ and $A(V, M)$. Towards this end, by a slight abuse of notation, we let the throughput at node i be denoted by $\Lambda_i(V)$, the system throughput be denoted by $\Lambda(V)$, and the availability by $A_i(V)$.

Let $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_f)$ be the maximum number of chargers allowed at each point due to limited space or power constraint. We further assume that all chargers are identical and have the same charging speed in this section, i.e., for mean charging time defined in (2), we have $t_i = t_j \ \forall i, j \in F$.

Let z_i be the average revenue per service at node $i \in I$. We further assume that there is a penalty of β_k dollars if there is a passenger loss, i.e., passenger finds no vehicle at departure point $k \in S$ and leaves the system. From Lemma 2, the penalty per unit time at node k is $\beta_k \alpha_k (1 - A_k(V))$, where α_k is the passenger arrival rate. Let c_i be the average cost for maintaining one charger at charging node $i \in F$ per unit time, which captures different rent and electricity rates at various places. Thus, the operating cost of chargers is $c_j v_j$ at charging point $j \in F$. The resulting optimization problem can be formulated as follows:

$$\max_{V \in \mathbb{N}^F} \sum_{i \in I} \Lambda_i(V) z_i - \sum_{k \in S} \beta_k \alpha_k (1 - A_k(V)) - \sum_{j \in F} c_j v_j \tag{10}$$

$$s.t.\ V \le \hat{V} \tag{11}$$

where the objective function is the revenue minus penalty due to loss of a passenger and the operating cost. We want to maximize it by controlling the number of chargers V at various charging points. The constraint means that the number of chargers at each charging point $i \in F$ is upper bounded by $\hat{v}_i$.

We now simplify the objective function. Similar to (6) in Lemma 1, we define the system throughput under a fixed M as

$$\Lambda(V) = \frac{G(M-1, V)}{G(M, V)}, \tag{12}$$

where G is the normalization constant introduced in (4). Therefore, following (5), the actual throughput of each node $i \in I$ is

$$\Lambda_i(V) = \lambda_i \Lambda(V).$$

Using (7), the above optimization problem can be rewritten as

$$\max_{V \in \mathbb{N}^F} \quad h(V) := \Lambda(V)\bar{Z} - \sum_{j \in F} c_j v_j - \sum_{k \in S} \beta_k \alpha_k \tag{13}$$

$$s.t.\ V \le \hat{V} \tag{14}$$

where $\bar{Z}$ is independent from V and defined as follows,

$$\bar{Z} = \sum_{i \in I} \lambda_i z_i + \sum_{k \in S} \lambda_k \beta_k \tag{15}$$

Let $e_j \in \{0, 1\}^f$ denote the unit vector with 1 along the jth dimension and 0 otherwise. In the following theorem, we show the concavity of objective function with respect to v_j, the number of chargers at charging point j, for all $j \in F$.

Theorem 9. *The following holds:*

1. *The map $v_j \mapsto \Lambda(V)$ is an increasing concave function for all $j \in F$.*
2. *The objective function $h(V)$ satisfies $h(V) + h(V + 2e_j) \le 2h(V + e_j)$ for all $j \in F$.*
3. *The map $v_j \mapsto A_i(V)$ is an increasing concave function for all $i, j \in F$.*

Proof Sketch: We first prove the first statement when the travel time distributions along the infinite server nodes are exponential. We then invoke Lemma 3 to conclude the first statement. This immediately yields the other two assertions since both $h(V)$ and $A_i(V)$ are linear with respect to $\Lambda(V)$. See Appendix C for more details.

This third part of the theorem points towards an interesting property of the network: adding chargers at any charging point will increase the system throughput and the availability of any departure point in the system. Therefore, service providers can

firstly allocate chargers to the charging points which can bring high system throughput increment at a low cost.

We now outline an algorithm, proposed in Section 4 of Shanthikumar and Yao (1988a), that computes an approximately optimal charger allocation in Algorithm 2. This algorithm is inspired by the marginal allocation algorithm of Fox (1966). The underlying idea for this algorithm is to identify the location where adding one more charger leads to the maximum increment in the profit. This process is continued until either the increment becomes negative or the upper bound is reached.

Algorithm 2: Charger Allocation Algorithm

Input : $h(V)$, $A(V)$, $\hat{V}$
Output: V^*

$k = 1$, $V^k = (1, 1, \ldots, 1)$, $\mathcal{F} = \{1, 2, \ldots, f\}$;
while $V^k \leq \hat{V}$ **do**
  if $v_j^k = \hat{v}_j$ **then**
    $\mathcal{F} \leftarrow \mathcal{F} - \{j\}$;
  end
  $m = \max_{j\in\mathcal{F}} h(V^k + e_j) - h(V^k)$;
  $j^* = \arg\max_{j\in\mathcal{F}} \; h(V^k + e_j) - h(V^k)$;
  if $m > 0$ **then**
    $V^{k+1} \leftarrow V^k + e_{j^*}$;
    $k \leftarrow k + 1$;
  else
    return V^k;
  end
end
return $\hat{V}$;

If there are only two finite-server queues in the system whose number of chargers may change, we can guarantee that the solution found by the above algorithm is optimal. For the general case, its optimality remains a conjecture (Shanthikumar & Yao, 1988a) and a proof of optimality is not available. In our computational experiments (see Sections 8.2.1 and 8.2.3), we see that this heuristic always returns optimal solutions in the simple cases we simulated.

Theorem 10. *If $|F| = 2$, then Algorithm 2 generates the optimal solution.*

Proof Sketch: Our proof follows the approach developed in Shanthikumar and Yao (1988a, Proposition 2, p. 339). We first show that our objective function h is concave and supermodular. The proof of convergence of the algorithm proposed in Shanthikumar and Yao (1988a) is for a fixed number of servers. On the other hand, we relax this constraint, since in our case we can put as many chargers as possible. A detailed proof is presented in Appendix D.

6. Charger selection

For electric vehicles, there are a large amount of chargers such as slow AC charger, DC fast charger, rapid charger and ultrafast charger. For example, the rapid chargers can charge the vehicle rather quickly; it can charge a vehicle from 20% charge to 80% charge within 30 min. On the other hand, slow AC chargers would require over 6–8 h to do the same. In this section, we define the fast and slow charger relatively: the charging time of fast charger is halved compared with slow chargers, regardless of types and techniques. Intuitively, one would conjecture that having one fast charger is better than two slow chargers. In this section, we identify the conditions under which it is beneficial to have two slow chargers as opposed to one fast charger to improve the overall throughput of the system. The key insight we get here is that the throughput of the system is dependent on the wait time for the vehicles as well as the charging time of the vehicles. Having one fast charger certainly reduces the charging time, but it can potentially increase the wait time if the *coefficient of variation* of the distribution of charging time of the vehicles is somewhat larger than a threshold. It should be noted that the installation cost of a fast charging infrastructure is significantly higher than installing multiple slow chargers, and we assume that they are the same in this section to simplify the analysis.

Suppose that the service provider has two possible options:

1. Install one fast charger with mean charging time t_0.
2. Install two slow chargers with mean charging time for each charger $2t_0$.

For ease of analysis, we discuss the charger selection problem by comparing individual queues under Poisson arrival. Let α_0 denote the Poisson arrival rate of vehicles that need to be charged. The mean charging rate (number of vehicles charged per unit time) for option 1 is $\mu_1(n) = \frac{1}{t_0}$. For option 2, the mean charging rate becomes

$$\mu_2(n) = \begin{cases} \dfrac{1}{2t_0}, & n = 1 \\ \dfrac{1}{t_0}, & n \geq 2 \end{cases} \tag{16}$$

where n is the number of vehicles at this charging point.

According to the model we assume in Section 2.2.4, we assume that upon arriving at a charging node, all the vehicles form an FCFS queue to wait for charging. Let $\gamma_1 = \frac{\alpha_0}{\mu_1}$ and $\gamma_2 = \frac{\alpha_0}{2\mu_2(1)}$ denote the utilization of the queue for two options respectively, then both options have the same utilization, $\gamma_1 = \gamma_2 = \alpha_0 t_0$.

Let D_1 and D_2 denote the average time delay of a vehicle at this node for both options (including waiting time and charging time). We show in the next section (see (24) and following discussion) that a smaller delay increases the system throughput of the closed queueing network. We want to find which option has a lower delay, thus it will have higher throughput, and as a consequence, a higher profit and better quality of service.

Intuitively, one fast charger outperforms two slow chargers because $\mu_1(1) > \mu_2(1)$. In the following lemma, we show that it is true for exponential charging time, however, it may not hold for some general charging time distribution, which is proved in the following theorem.

Lemma 11. *If charging time distribution is exponential, then $D_1 < D_2$, $\forall\ \gamma \in (0, 1)$,*

Proof. For $M/M/1$ queue with arrival rate α_0, service rate μ_1 and utilization $\gamma_1 = \frac{\alpha_0}{\mu_1}$, the average waiting time in queue is $w_1 = \frac{\gamma_1}{\mu_1(1-\gamma_1)}$ from Smith (2018, p.82). Therefore, the average delay for option 1 is $D_1 = w_1 + t_0 = \frac{t_0}{1-\gamma_1}$.

For $M/M/2$ queue with arrival rate α_0, service rate $\mu_2(n)$ and utilization $\gamma_2 = \frac{\alpha_0}{2\mu_2(1)}$, the average waiting time in queue is $w_2 = \frac{\gamma_2^2}{\mu_2(1)(1-\gamma_2^2)}$, Smith (2018, p.87). Thus the average delay for option 2 is $D_2 = w_2 + 2t_0 = \frac{2t_0}{1-\gamma_2^2}$.

As $\gamma_1 = \gamma_2 \in (0, 1)$, we can conclude that $D_1 < D_2$, $\forall \gamma \in (0, 1)$. □

Let $c^2 = \frac{Variance}{Mean^2}$ denote the squared coefficient of variance of charging time, which measures the dispersion of the charging time distribution.

Theorem 12. *For any $\gamma \in (0, 1)$, there exists a charging time distribution such that $D_1' > D_2'$ for all $c^2 > 1 + \frac{2}{\gamma}$.*

Proof. We prove this theorem by providing an extreme case of charging time distribution. For option 1, called $M/T_1/1$ queue, the charging time follows the distribution

$$T_1 = \begin{cases} \exp(p_0/t_0) \text{ w.p. } p_0 \\ 0 \text{ w.p. } 1-p_0 \end{cases}$$

where $\exp(p_0/t_0)$ denotes the exponential distribution with mean t_0/p_0. Then the mean of T_1 is t_0 and squared coefficient of variance is $c_1^2 = \frac{2}{p_0} - 1$. From P-K formula, we have $D_1' = \frac{t_0}{p_0}\frac{\gamma_1}{1-\gamma_1} + t_0$.

For option 2, $M/T_2/2$ queue, its charging time for each charger follows the same distribution of option 1 with double mean, i.e.

$$T_2 = \begin{cases} \exp(p_0/2t_0) \text{ w.p. } p_0 \\ 0 \text{ w.p. } 1-p_0 \end{cases}$$

So the mean of T_2 is $2t_0$ and squared coefficient of variance is $c_2^2 = \frac{2}{p_0} - 1$, same as c_1^2.

In the following, we calculated the average delay $D_2' = \omega_2' + 2t_0$ for option 2: $M/T_2/2$ queue, where ω_2' is the averaging waiting time and $2t_0$ is the average service time.

First, we show that ω_2' is equivalent to the average waiting time ω_3' in $M/M/2$ system with arrival rate $\lambda_3' = p_0\alpha_0$ and unit service rate $\mu_3'(1) = \frac{p_0}{2t_0}$. Since the scheduling discipline is independent of the service time, the waiting time experienced by non-zero jobs and zero-sized jobs in T_2 is the same. Further, to find the waiting time of non-zero jobs, we can ignore the zero-sized jobs, since adding/removing zero-sized jobs will have no impact on the waiting time of non-zero jobs. According to the definition of T_2, this $M/T_2/2$ queue will have the same behavior with such $M/M/2$ queue when considering non-zero jobs, with arrival rate $\lambda_3' = p_0\alpha_0$ and unit service rate $\mu_3'(1) = \frac{p_0}{2t_0}$. Therefore we can conclude that $\omega_2' = \omega_3'$.

Second, we calculate ω_3' similar to $M/M/2$ queue in Lemma 11. As the utilization is defined as $\gamma_3 = \frac{\lambda_3'}{2\mu_3'(1)} = \alpha_0 t_0$, the average waiting time in queue is $w_3' = \frac{\gamma_3^2}{\mu_3'(1)(1-\gamma_3^2)}$, Smith (2018, p.87). As $\gamma_2 = \gamma_3 = \alpha_0 t_0$, we have $w_3' = \frac{2t_0\gamma_2^2}{p_0(1-\gamma_2^2)}$.

Finally, we have the average delay $D_2' = w_2' + 2t_0 = \frac{2t_0\gamma_2^2}{p_0(1-\gamma_2^2)} + 2t_0$.

As $\gamma_1 = \gamma_2$, we have $D_1' > D_2'$ is equivalent to $\frac{\gamma}{1+\gamma} > p_0 > 0$. As $c_1^2 = c_2^2 = \frac{2}{p_0} - 1$, we have $c^2 > 1 + \frac{2}{\gamma}$. □

The above theorem indicates that two slow chargers can result in a lower overall delay than one fast charger, especially when the variance in charging time is relatively larger than the mean charging time. As stated previously, this happens because the large variance in charging time leads to a long waiting time, which reduces the waiting time by adding one more charger. Although the average charging time is doubled due to slow chargers, the decrease in waiting time is more significant than the increase of charging time, thus the total delay may decrease with two slow chargers.

Admittedly, the distribution constructed in the proof of Theorem 12 is not representative of the actual charging time distribution. Nonetheless, we have found through simulations that for various distributions of charging time, there is a distribution dependent threshold for c^2, beyond which two slow chargers have lower delay (wait time plus charging time) than one fast charger. Indeed, in Section 8.3, we show numerically the case for gamma distributed charging times and Inverse Gaussian distributed charging times.

7. Computational algorithms

In this section, we introduce some efficient algorithms for performance analysis, especially for large-scale networks.

7.1. Convolution algorithm

In order to compute the stationary state probability, the normalizing constant $G(M)$ is required as stated in (4). Explicitly, $G(M)$ has the following expression.

$$G(M) = \sum_{n_1+\cdots+n_N=M} \prod_{i=1}^{N} \frac{\lambda_i^{n_i}}{\prod_{k=1}^{n_i} u_i(k)} \tag{17}$$

Direct computation of $G(M)$ as a summation over all possible states, which has a cardinality of $\binom{N+M-1}{N-1}$, takes an exponential time to compute. However, we can use a convolution algorithm, which significantly reduces the complexity by developing an iterative algorithm.

Following the definition in Section 3, we assume that there are M vehicles and N nodes in the system. For the case where n_i vehicles at node $i \in N$, we define

$$k_i(n_i) = \frac{\lambda_i^{n_i}}{\prod_{k=1}^{n_i} u_i(k)} \tag{18}$$

Let $G_N(M)$ denote the normalizing constant of network with M vehicles and N nodes, then we have (Buzen, 1973)

$$G(M) := G_N(M) = k_1 * k_2 * \ldots * k_N(M) \tag{19}$$

where the convolution $k_1 * k_2(m)$ of two functions k_1 and k_2 is defined by

$$k_1 * k_2(m) = \sum_{i=0}^{m} k_1(i)k_2(m-i), \ m \geq 0 \tag{20}$$

We can write the recursive relation in another way, for each $j = 1, 2, \ldots, N$, we have

$$G_j(m) = k_j * G_{j-1}(m), \ 0 \leq m \leq M \tag{21}$$

Therefore, we can get the stationary probability distribution P from (4), the throughput $\Lambda(M)$ from (6), and the marginal probability $p_i(n_i)$ for each node from (A.1).

7.2. Mean value analysis

As M becomes larger, the convolution still takes a large amount of time, especially when the network has several load-dependent queues. If we are only interested in the average performance metrics, we can utilize the Mean-Value analysis of closed queueing networks (Akyildiz & Bolch, 1988; Reiser & Lavenberg, 1980) in order to compute the outcome easier.

Let $D_i(M)$ denote the average system time (waiting time and charging time) of a passenger at node $i \in \mathcal{N}$. According to three different scheduling schemes (infinite, single, finite server), we have

$$D_i(M) = \begin{cases} \dfrac{1}{u_i(1)}, & i \in I \\ \dfrac{1 + L_i(M-1)}{u_i(1)}, & i \in S \\ \dfrac{1 + L_i(M-1) + s_i(M-1)}{v_i u_i(1)}, & i \in F, \end{cases} \tag{22}$$

where $L_i(M)$ is the average number of vehicles (including the one in service) at node i when fleet size is M, and v_i is the number of chargers at finite-server nodes F.

We further define $s_i(M-1)$ as follows, which is the average number of idle chargers at node $i \in F$, then we have

$$s_i(M-1) = \sum_{n_i=1}^{v_i-1} (v_i - n_i) p_i(n_i - 1, M-1) \tag{23}$$

where $p_i(n_i, M)$ is the marginal probability of n_i vehicles at node $i \in F$ when the fleet size is M.

Therefore, we have the system throughput as follows where λ_i is defined in (3).

$$\Lambda(M) = \frac{M}{\sum_{i=1}^{N} \lambda_i D_i(M)} \tag{24}$$

Applying Little's Law, we can compute the queue length $L_i(M)$ by iteration without computing the normalization constant $G(M)$ as follows:

$$L_i(M) = \lambda_i \Lambda(M) D_i(M) \tag{25}$$

Therefore, we can iterate over M to compute the throughput $\Lambda(M)$ from (24), where $D_i(M)$ only requires the information of last stage $L_i(M-1)$ and $s_i(M-1)$. One more expression $p_i(n_i, M)$ in (23) needs to be specified.

In order to compute the marginal distribution $p_i(n_i, M)$ in (23), for $i \in F$, we have to run another iteration with respect to n_i.

For $n_i = 1, 2, \ldots, M$, from local balance we have

$$p_i(n_i, M) = \frac{\lambda_i \Lambda(M) p_i(n_i - 1, M-1)}{n_i u_i(1)} \tag{26}$$

and

$$p_i(0, M) = 1 - \frac{1}{v_i}\left(\frac{\lambda_i \Lambda(M)}{u_i(1)} + \sum_{n_i=1}^{v_i-1}(v_i - n_i) p_i(n_i, M)\right) \tag{27}$$

In this way, we can compute the throughput and availability faster than the convolution algorithm.

7.3. General passenger inter-arrival time approximation

In practice, passengers may not follow a Poisson arrival, that is, the distribution of passenger inter-arrival time may not follow the exponential distribution. In this subsection, we show that we can approximate the passenger arrival process through a modification of the system equation in the Mean Value Analysis algorithm.

If the inter-arrival time is exponentially distributed, then it enjoys the memoryless property: We do not need to consider the remaining service time for vehicles at $i \in S$. If the inter-arrival time follows a general distribution with mean $\frac{1}{u_i(1)}$ and variance σ^2, we need to consider the remaining service time. Let $c_\tau^2 = \sigma^2 u_i(1)^2$ denote the squared coefficient of variance of the inter-arrival time. Use $\gamma_i = \frac{\lambda_i}{u_i(1)}$ denote the relative utilization and the true utilization $\rho(M-1) = \gamma_i \Lambda(M-1)$ for the case of $M-1$ vehicles in the system. If the service time of single server nodes (SS) becomes general, then for each $i \in S$, the system time can be approximated by Curry and Feldman (2010, p. 253) and Smith (2018, p. 291)

$$D_i(M) = \frac{1}{u_i(1)}\left(1 + L_i(M-1) - \rho(M-1) + \rho(M-1)\frac{1+c_\tau^2}{2}\right), \quad \forall i \in S \tag{28}$$

By replacing the second item in (22) with the above equation, we can approximate the stationary distribution if the inter-arrival time has a general distribution. If $c_\tau^2 = 1$ holds, the above equation goes back to the previous system time under exponential service time distribution in (22). Similar approximating methods could be extended to the case with a general charging time distribution, where c_τ^2 is the squared coefficient of variance for charging time (Gupta et al., 2010).

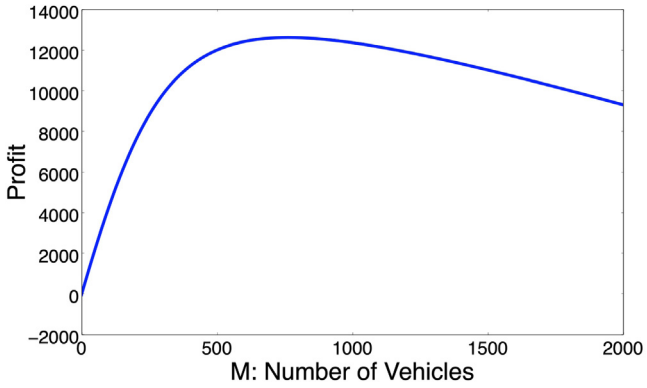


Fig. 3. Profit as a function of fleet size under 60 stations.

8. Numerical simulation

In this section, we run some large-scale simulations to validate our results. The first part develops a large-scale symmetric network and studies the asymptotic properties proved before. The second part focuses on the charger allocation on an asymmetric network capturing different characteristics of downtown and suburban areas. The third part shows the effect of one fast charger and two slow chargers.

8.1. A symmetric network with 60 nodes

Firstly, we consider one symmetric network with 60 stations with $p_{ij} = \frac{1}{59}$, i.e. after departure from one station, customers choose their destination equally between other stations. We further assume that one-third of EVs arriving at node j decide to charge, while others choose to go to departure point directly without charging, i.e., $\bar{p}_i = \frac{1}{3}$, $\forall i \in F$.

In this network, we have 60 single server nodes (departure), 60 finite server nodes (charging) and 3540 infinite server nodes (traveling). From (3), we have the relative throughput as $\lambda_i = \frac{1}{420}$ for $i \in F$, $\lambda_i = \frac{1}{140}$ for $i \in S$ and $\lambda_i = \frac{1}{8260}$ for $i \in I$. The arrival rate of customers at each departure point $i \in S$ is $\alpha_i = 10$ person per hour. The average time of traveling is $T_{jl} = \frac{1}{3}$ hour per service (service rate: 3 per hour), which follows a general distribution. The average time of charging is $t_i = 0.5$ hour per charger $i \in F$ (service rate: 2 per hour) and there are $v_i = 2$ chargers at each charging station.

8.1.1. Fleet sizing

We assume the average revenue per service is $z_i = \$30$ and the operating cost per vehicle is \$4 per hour $g(M) = 4M$. Let $\epsilon_i = 20\%$ for $i \in S$, which means there is a minimal requirement of 80% availability at each departure point.

Under above assumptions, the simulations at Figs. 3 and 4 validate the concavity proved in Lemma 5 and Theorem 7 it also shows that the optimal fleet size is 763, with availability 87.2%.

If the number of chargers is one at each station, the availability requirement is not satisfied (only 54.47%), thus we start from two chargers per station. As shown in Fig. 5, the optimal fleet size increases as the availability requirements increases. On the other hand, the optimal fleet size decreases as more chargers are provided at each charging point, as shown in Fig. 6, and the curve remains stable when chargers are relatively high. This is because more chargers will decrease the waiting time before charging and provide more availability as less time is spent at charging points.

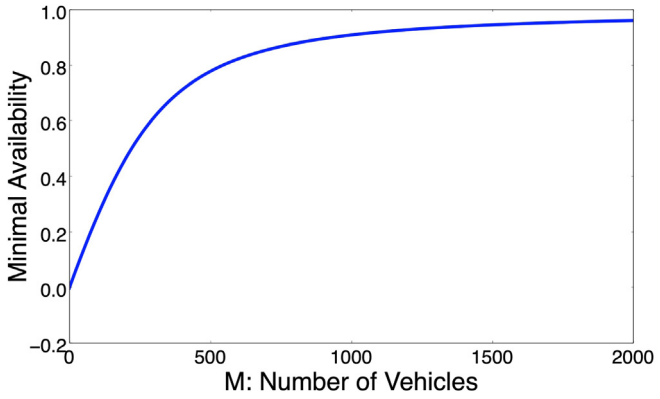


Fig. 4. Availability at departure points as a function of fleet size under 60 stations.

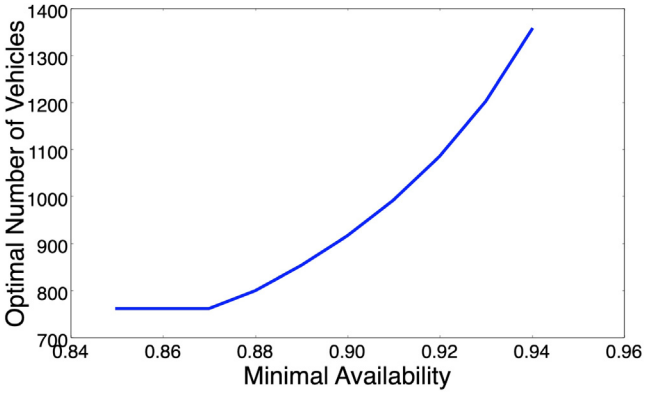


Fig. 5. Optimal fleet size increases dramatically with higher availability requirements.

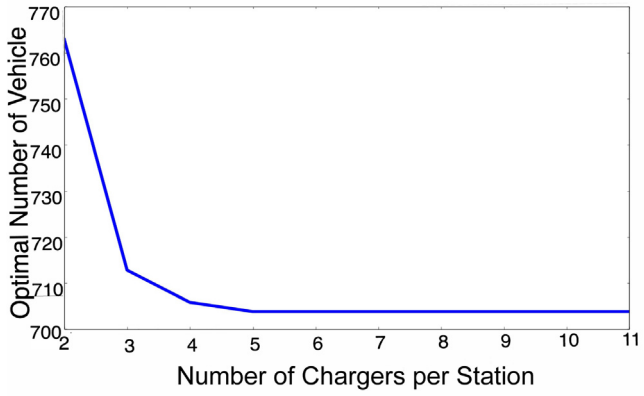


Fig. 6. Optimal fleet size decreases dramatically as number of chargers increases and it keeps flat when chargers are relatively large, with same availability requirement 80%.

8.1.2. Charger allocation

We now study the effect of number of chargers on the throughput, availability, and profit numerically for the example introduced above. We fix the fleet size to be $M = 763$. Other parameters remain the same as in the last subsection such as transition probability, charging probability, arrival rate, charging time, travel time, and revenue per service. Suppose that the operating cost per charger is $c_j = 2$ for all $j \in F$. The penalty for one passenger loss is $\beta_k = 1$ for all $k \in S$.

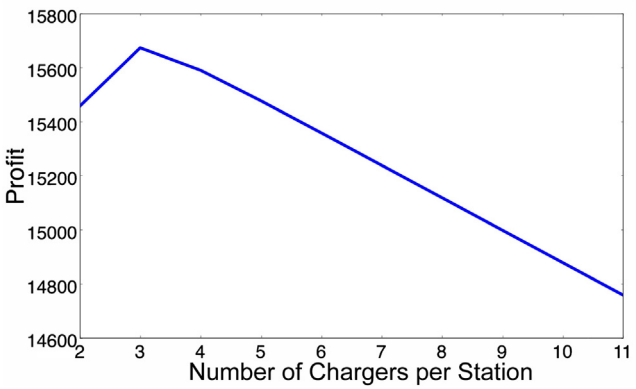


Fig. 7. Profit is a concave function of number of charger per station under charger operating costs and fixed fleet size.

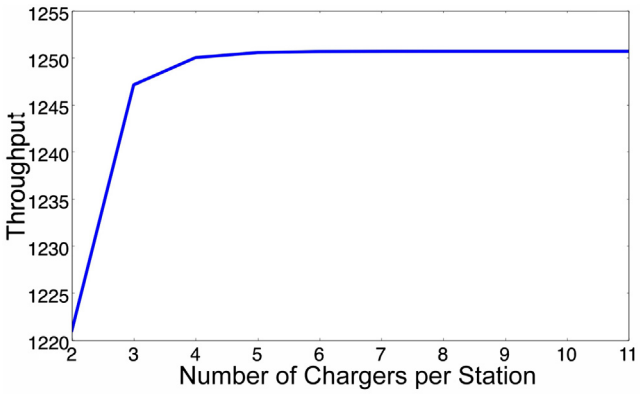


Fig. 8. Throughput increases as the increase of number of charger.

As established in Theorem 9, the profit is a concave function of the number of chargers. This can be seen in Fig. 7. Moreover, we observe that the profit maximizing point is to have 3 chargers per station for the numerical example considered here.

Fig. 8 depicts the system throughput as a function of the number of chargers at each node. We observe that the system throughput increases as we increase the number of chargers. This is because the vehicles spend less time waiting for charging at charging points. When the number of chargers at each node is larger than a threshold, then the system throughput does not increase as we increase the number of chargers. For the current numerical example, this threshold is 4 charger per station.

We now focus on the availability at one departure point as shown in Fig. 9. As more chargers become available, electric vehicles spend less time at charging points and more time at departure points. Therefore, the availability increases at single server queues.

Leveraging the Mean Value Analysis in Section 7.2, the average queue length of each node (average number of vehicles at this node) is depicted in Fig. 10. As the number of chargers increases, the delay at charging points (Finite Server nodes) decreases because of less waiting time, and the vehicles move to the departure points (Single Server nodes), which increases their availability.

8.2. Asymmetric network capturing the relation between downtown and suburban areas

In this subsection, we allocate chargers when the network is not symmetric. We use a three-station network in which the

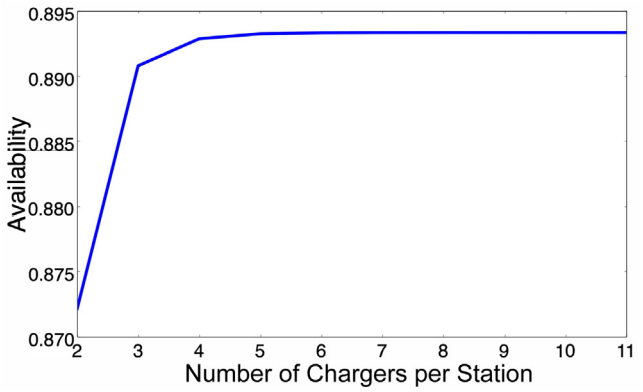


Fig. 9. Availability of vehicle increases if more chargers are installed at charging points.

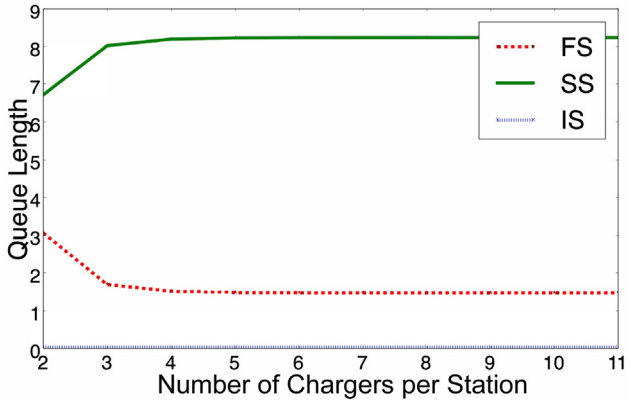


Fig. 10. Average number of vehicles at finite server nodes (FS:charging), single server nodes (SS:departure) and infinite server nodes (IS;travel) changes, with respect to number of chargers.

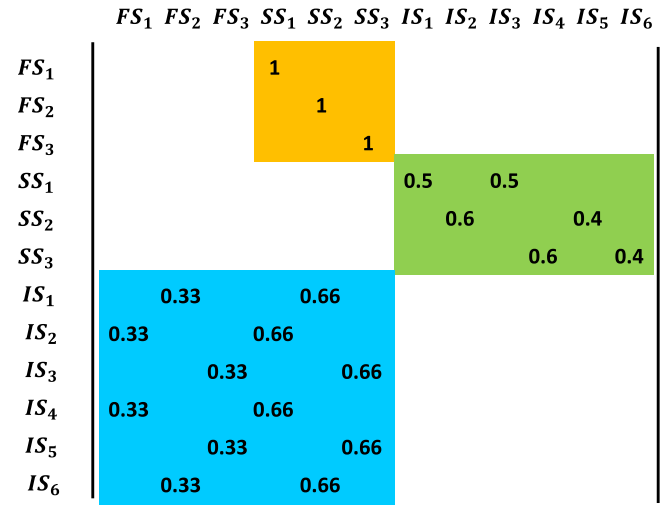


Fig. 11. Routing matrix r_{ij} between nodes under asymmetric three-station network, where station 1 denotes downtown and station 2&3 denote suburban areas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Charger allocation algorithm.

Step	V	Profit	Revenue	Cost	Penalty
1	(1,1,1)	458.16	478.25	8	12.09
	(1,2,1)	457.53	479.56	10	12.03
2	(2,1,1)	533.58	554.79	12	9.21
	(3,1,1)	530.05	555.23	16	9.18
3	(2,2,1)	553.46	575.89	14	8.43
	(3,2,1)	549.53	575.93	18	8.40
4	(2,2,2)	766.58	783.21	16	0.63
	(2,3,2)	766.98	785.55	18	0.57
5*	(3,2,2)	769.61	790.00	20	0.39
	(3,3,2)	769.52	791.85	22	0.33
	(4,2,2)	766.15	790.51	24	0.36

first station denotes the downtown and the second and the third station denote suburban areas.

As shown in Fig. 11, we have the routing matrix r_{ij} from (1), where the orange part denotes that every vehicle will enter the departure point after finishing charging. The green part denotes the asymmetric transition probability p_{ij} between stations, i.e., passengers departing from station 1 (downtown) will choose their destination equally between station 2 and 3 (suburb), while passengers departing from station 2 or 3 (suburb) will choose station 1 (downtown) as their destination with probability 60% while station 3 or 2 (suburb) with probability 40%. The blue part denotes that statistically one-third of EVs arriving at node $i \in F$ decide to be charged, while others choose to go to departure point directly without charging, i.e., $\bar{p}_i = \frac{1}{3}$, $\forall i \in F$. Other parts in this routing matrix are zero.

According to global balance (3), we have the relative throughput as follows, where the order of nodes is the same as the routing matrix.

$$\lambda = \left[\frac{3}{56}, \frac{5}{112}, \frac{5}{112}, \frac{9}{56}, \frac{15}{112}, \frac{15}{112}, \frac{9}{112}, \frac{9}{112}, \frac{9}{112}, \frac{9}{112}, \frac{3}{56}, \frac{3}{56} \right] \tag{29}$$

8.2.1. Charger allocation

With a fixed number of vehicles 40 in the system, we follow the Algorithm 2 in order to find the optimal charger allocation in this network.

Distinguishing the difference of rent between downtown and suburb, we assume that the operating cost per charger is \$2 per hour at station 2 and 3 (suburb) while \$4 per hour at station 1 (downtown). We further assume the average revenue per service is \$30 and the penalty is \$ 1 per loss.

Following the Charger Allocation Algorithm (Algorithm 2), we can compute the allocation solution without listing all the candidates. As shown in Table 1, the algorithm reduces the candidate size and computing complexity by leveraging the concavity property. After step 5, the algorithm terminates and we can claim that (3, 2, 2) is the optimal solution in our setting, i.e., the number of chargers at (FS1,FS2,FS3) is (3, 2, 2), without analyzing any other candidates.

If the constraint is active, i.e., V is upper bounded by $\hat{V}$ as in (14), this algorithm can still work. For example, if $\hat{V} = (2, 5, 5)$ is the upper bound of V, then the algorithm terminates at (2, 3, 3), indicating this is an approximate solution to the optimization under this constraint.

8.2.2. Convolution

Under the fleet size $M = 40$ and $V = (3, 2, 2)$, we can use convolution algorithm in Section 7.1 and (A.1) to compute the distribution of electric vehicles, through the marginal distribution of each node, i.e., the probability of n_i vehicles at node $i \in \mathcal{N}$ in the closed queueing system.

For example, the Figs. 12 and 13 present the marginal distribution in the departure and charging points of suburban areas. The first figure shows that there is no vehicle waiting at SS2

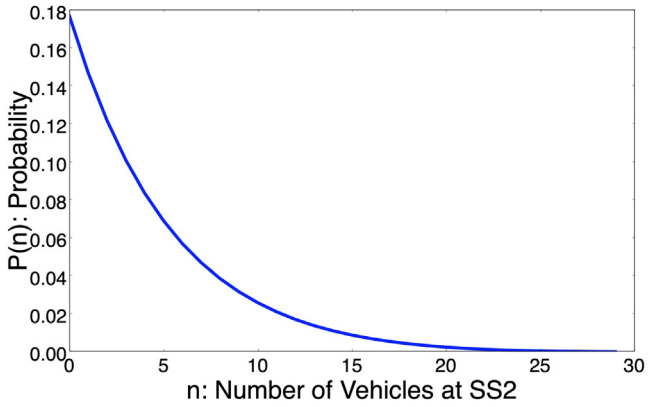


Fig. 12. Marginal Distribution of vehicles at SS2 (suburb-departure) under $M = 40$ and $v = \{3, 2, 2\}$.

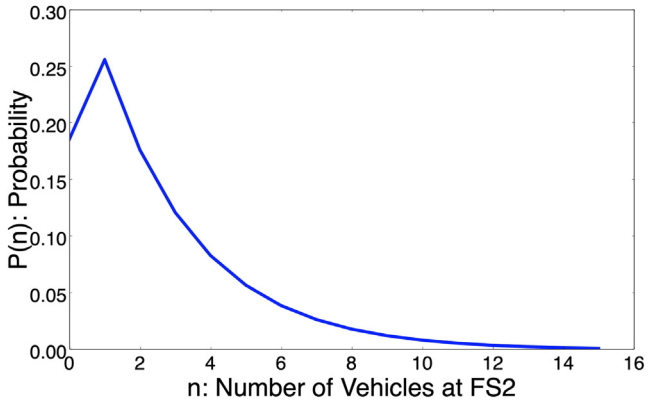


Fig. 13. Marginal Distribution of vehicles at FS2 (suburban charging station) under $M = 40$ and $v = \{3, 2, 2\}$.

with a probability 18%, which means that the availability at SS2 is 82%. The second figure shows that there is no vehicle charging at FS2 with probability 18%, which indicates that the utilization of charging infrastructure is high.

8.2.3. Large-scale charger allocation

We assess the charger allocation algorithm with a larger-scale network, where the number of stations increases from 3 to 10 and the number of vehicles is 50. Station 1 is the downtown and station 2–10 are suburb. Vehicles departing from downtown will choose other stations equally with probability 1/9, while passengers departing from suburb choose station 1 (downtown) with probability 20% and other stations equally with probability 10%. The operating cost per station is \$4 per hour at station 1, \$3 per hour at station 2–5 and \$2 per hour at station 6–10. Other parameters are the same as Section 8.2.1.

As shown in Fig. 14, the profit at initial stage with $V = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$ is \$986 per hour. After 16 iterations in Algorithm 2, the profit increases to \$1634 per hour with $V^* = \{3, 2, 2, 2, 2, 3, 3, 3, 3, 3\}$. We further compare it with 5^{10} candidates of charger allocation, where the number of chargers varies from 1 to 5 at each station. V^* turns to be the optimal solution in this problem, i.e. we do not find the situation when Algorithm 2 achieves a sub-optimal solution.

8.3. One fast vs. Two slow

In this section, we show the result in Section 6 about one fast server vs. two slow servers. We test them in a simple closed

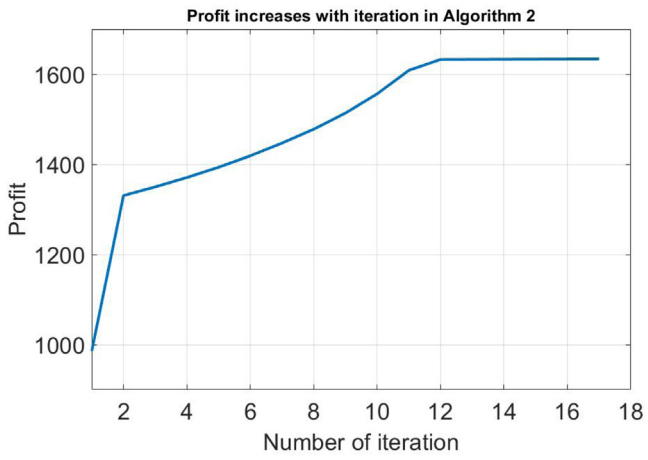


Fig. 14. Profit increases with iteration in Algorithm 2, the number of chargers at various stations increases from $V = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$ to $V = \{3, 2, 2, 2, 2, 3, 3, 3, 3, 3\}$.

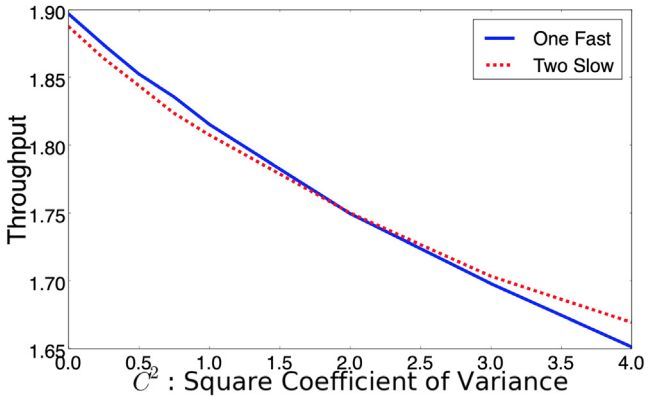


Fig. 15. Throughput comparison of one fast charger and two slow chargers under gamma distributed charging time in a closed queueing system.

queuing network of two queues, where the output of one queue is the input of another queue. The first queue is a single server queue with exponential service time, whose mean service time is 1/2. The second queue has two choices, as shown in Section 6 with $t_0 = 1/2$. i.e., one fast charger with mean time 1/2 vs. two slow charges with mean time 1 We compare the system throughput of two choices.

Assume that there are only 10 vehicles in this closed queueing network, we use Monte-Carlo simulation to find the system throughput under gamma distribution.

As shown in Fig. 15, when squared coefficient of variance c^2 is small, $D_1 < D_2$ and thus one fast server has larger throughput. As c^2 increases, the gap between two choices closes. When c^2 becomes large, two slow servers outperform one fast server, i.e., $D_1 > D_2$ and the throughput under two slow servers is larger.

If we keep increasing the number of chargers, the curve will be flatter, where the mean charging time for each charger is $n/2$ and n is the number of slow chargers. If the number of chargers exceeds the number of vehicles in the system, the throughput does not change with respect to the squared coefficient of variance, as shown in Lemma 4.

As shown in Fig. 16, the threshold of c^2 where one server outperform five servers is around 4, which is larger than 1.9, the threshold where one server outperform two servers in Fig. 15. When number of server is 10, which is the same as number of vehicles in the system, the throughput does not change with

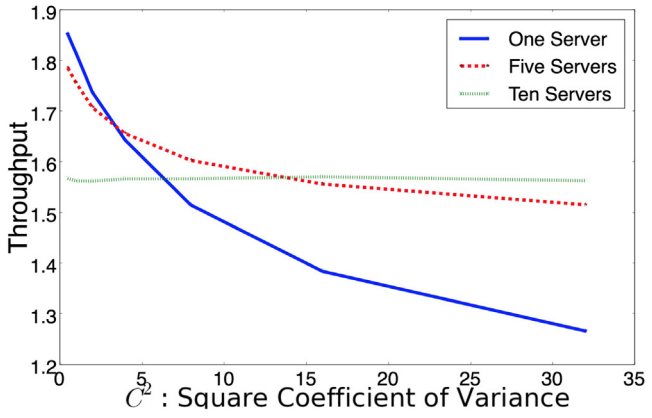


Fig. 16. Throughput comparison of one, five and ten servers under Gamma distributed charging time.

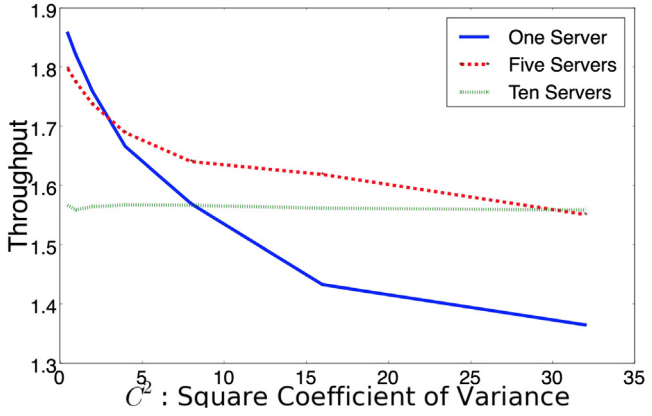


Fig. 17. Throughput comparison of one fast server and two slow servers under Inverse Gaussian distributed charging time.

respect to the variance of charging time. Fig. 17 shows that similar results hold for Inverse Gaussian distributed charging time and the threshold for c^2 is distribution dependent.

8.4. SoC-dependent charging

We use Monte Carlo simulation to evaluate the impact of SoC independent charging time assumption. Similar to simulation in Section 8.2, we use a small closed queueing network with 3 stations. The routing between stations is symmetric and the travel time is 0.3 between station A and B, 0.4 between B and C, 0.5 between A and C. After finishing one ride at IS node, EV will be charged at the following charging point at FS node, with probability 30%, 40%, 50%, respectively. The associated mean charging time has two options (i) Soc-consumption dependent, i.e., 0.3, 0.4, 0.5 for three types of travels. In this case, larger consumption of energy on ride will incur a larger probability for charging and a larger charging time. (ii) Soc-consumption independent, i.e. 0.357 for charging point FS1, 0.425 for FS2, 0.455 for FS3. In this case, the mean time is the average of different charging time weighted by their throughput (demand). We compare these two options on cases with different number of chargers. As shown in Fig. 18, the approximation error is less than 1 percent and we believe this error is acceptable. The seven charge vectors are $V_1 = [1, 1, 1]$, $V_2 = [1, 1, 2]$, $V_3 = [2, 1, 2]$, $V_4 = [2, 2, 2]$, $V_5 = [2, 2, 3]$, $V_6 = [3, 2, 3]$, $V_7 = [3, 3, 3]$.

8.5. Real distribution for charging time

We first download the real data for charging time at Adaptive Charging Network (ACN-Data) 2021 and plot it in Fig. 19. In this dataset, charge time is calculated as the difference between 'connectionTime' and 'disconnectTime', regardless of whether the charge is complete or not. The mean charge time is 2.92 h (10 508 s) with maximum 13.84 h and minimum 0.03 h.

Although the charging time distribution is not exactly the same as exponential distribution, we use a simulation to show that the approximation error is acceptable, if we approximate the real charging time distribution with an exponential distribution with mean 2.92. Similar to the simulation in Section 8.3, we use a simple closed queueing network with 10 vehicles, where the output of one queue is the input of another queue. The first queue is a single server queue (SS) with exponential service time (mean 0.5), which indicates vehicle arrive at the charging point according to the Poisson process with a mean arrival rate 2 per hour. The second queue is a finite server queue (FS) with number of chargers spanning from 1 to 10. The charge time in each charger has two cases, (i) it follows the exponential distribution with mean 2.92, (ii) it is randomly drawn from the real charging time dataset mentioned above. The throughput is calculated through Monte Carlo simulation and the approximation error is defined as the difference between two throughputs over the real throughput. As shown in Fig. 20, the approximation error is less than 2 percent when the number of chargers increases from 1 to 10. We believe that this error is acceptable. We also note that the same approximation is used in Jung et al. (2014) and Yang et al. (2017).

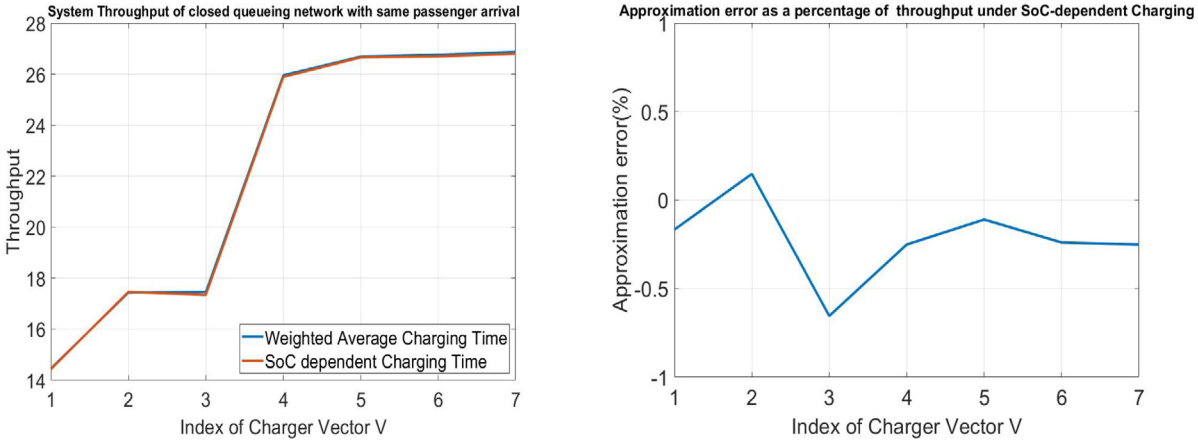
9. Conclusion

In this paper, we developed a closed queueing model for modeling a fleet of electric vehicles providing transportation service in a city. We considered the fleet sizing problem to maximize the profit of the system and the number of charges allocated within each charging station to maximize the total operational cost. We proved that the two problems lead to convex integer optimization problems. We developed a greedy algorithm for charger allocation and established its optimality if there are only two charging stations. When the variance of charging time becomes larger than the mean charging time, we showed using a stylized example that two slow chargers outperform one fast charger in terms of the total delay (waiting time plus the charging time). We further developed an approximation method for general passenger inter-arrival time distributions and the mean value analysis algorithm is provided for the performance analysis of the overall system.

Through this analysis, we gained many insights about fleet sizing, charger allocation, and charger selection.

1. As shown in Fig. 6, the optimal fleet size can be reduced by adding more chargers.
2. Theorem 9 shows that adding chargers at any charging point will increase the system throughput and the availability of any departure point in the system.
3. We posit that chargers should be allocated to the charging points that can bring high system throughput increment with a low cost, which is usually the areas with high visit ratios. This idea is inspired by the marginal allocation scheme studied in operations research (Fox, 1966).
4. Fast chargers may be replaced by multiple slow chargers, if the standard variance of charging time is relatively large compared with mean charging time. We show this insight to be useful through a numerical simulation Section 8.3 (see Fig. 15).

Future research will address the rebalancing policy, more general charging time distributions and state-dependent routing strategies.



(a) Throughput (b) Approximation Error

Fig. 18. Throughput comparison and approximation error between SoC-dependent charge time and weighted average charge time.

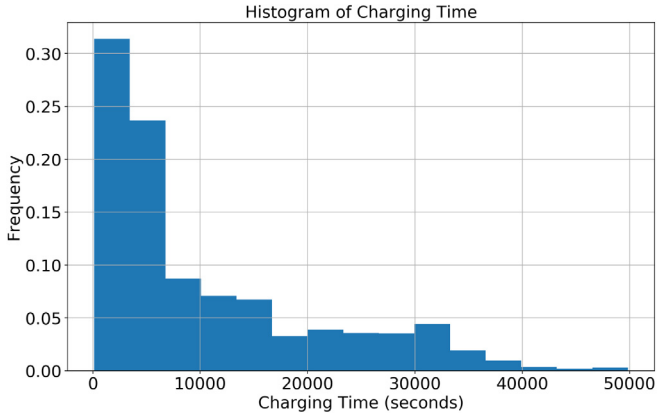


Fig. 19. Histogram of real charging time based on ACN-Data 2021 dataset (https://ev.caltech.edu/dataset).

CRediT authorship contribution statement

Yuntian Deng: Conceptualization, Methodology, Writing – original draft. **Abhishek Gupta:** Writing – review & editing. **Ness B. Shroff:** Supervision, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

This research is supported by Ford Motor Company (US) under the University Alliance Project.

Appendix A. Proof of Lemma 1

1. For a product form closed queueing network with M vehicles, Eqs. (5) and (6) are given by Serfozo (2012, p.27) where λ_i is the visit ratio defined in (3) and $G(M)$ is the normalization factor defined in (4) and (17).

2. The idea comes from George and Xia (2011) and we rewrite the equations with our notations as follows. The marginal distribution, i.e., the probability of n_i vehicles at node $i \in \mathcal{N}$ is:

$$p_i(n_i) = \frac{\lambda_i^{n_i}}{\prod_{k=1}^{n_i} u_i(k)} \frac{G_i(M-n_i)}{G(M)} \tag{A.1}$$

where $G_i(M-n_i)$ is the normalizing constant when node i is removed and only $M-n_i$ vehicles remains in the system by Lavenberg (1983, p.128.). For nodes $i \in S$, the special case when the node is a single server node (SS), we can compute the probability without computing $G_i(M-n_i)$ We define the relative utilization for single server node $i \in S$ as

$$\gamma_i = \frac{\lambda_i}{u_i(1)} = \frac{\lambda_i}{\alpha_i}$$

Then the probability of n_i vehicles at departure point $i \in S$ can be simplified by Lavenberg (1983, p.128) as follows.

$$p_i(n_i) = \frac{\gamma_i^{n_i}[G(M-n_i) - \gamma_i G(M-n_i-1)]}{G(M)}$$

The availability is defined as the stationary state probability that node i has at least one vehicle and has the expression as follows.

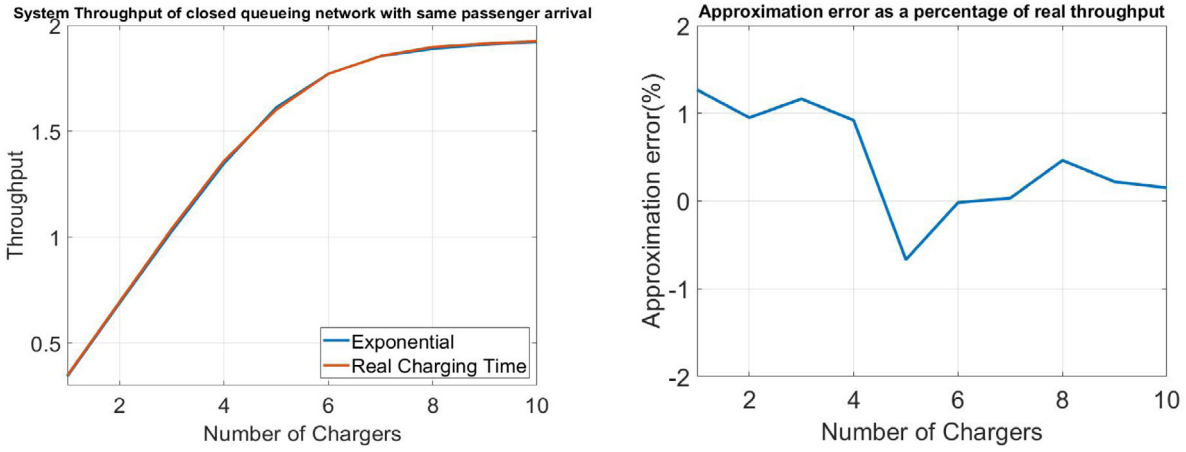
$$A_i(M) = 1 - p_i(0) = \gamma_i \Lambda(M) = \frac{\lambda_i}{\alpha_i} \Lambda(M) \tag{A.2}$$

Appendix B. Proof of Lemma 5

As shown in Shanthikumar and Yao (1988b, Theorem 1), in a closed Jackson network (exponential service time), the system throughput $\Lambda(M)$, is nondecreasing concave with job population M, if the service rate $u_i(n_i)$ is nondecreasing concave with local queue length n_i, $\forall i \in \mathcal{N}$.

We first consider the situation that the service time of the infinite server (IS) follows exponential distribution in our problem, then our network falls into the Jackson network. In our setting, $u_i(n_i)$ defined in (2) is constant (Single Server), linear (Infinite Server) and nondecreasing concave (Finite Server) with respect to n_i, i.e., all the service rates satisfy the nondecreasing concavity condition. Therefore, the system throughput $\Lambda(M)$ defined in (6) is also non-decreasing concave with M.

Secondly, if we change the service distribution of infinite server queues (IS) into a general distribution with the same mean,



(a) Throughput (b) Approximation Error

Fig. 20. Throughput comparison and approximation error between real charge time and exponential distributed charge time.

by Lemma 3, the throughput $\Lambda(M)$ does not change and the non-decreasing concave property remains.

Finally, applying (6), we have

$$\sum_{i\in I} z_i \Lambda_i(M) = \Lambda(M) \sum_{i\in I} z_i \lambda_i$$

where both z_i and λ_i are independent of M, which means the first part of objective function is concave.

As $g(M)$ is convex, then the second part of $f(M)$ is concave. As a conclusion, $f(M)$ is concave since the sum of two concave functions yields a concave function.

Appendix C. Proof of Theorem 9

Let us first prove the first statement when the travel time distributions along the infinite server nodes are exponential. We then invoke Lemma 3 to conclude the first statement. This immediately yields the other two assertions.

1. Firstly, assume that all the service time are exponentially distributed in our problem. From (2), we find that for all queues $i \in \mathcal{N}$, the service time $u_i(n_i)$ is increasing concave with n_i. From Shanthikumar and Yao (1987, Theorem 1), in such closed queueing network, if there is a finite server node $j \in F$ with v_j servers, i.e. service rate $u_j(n_j) = u_j(1)min\{n_j, v_j\}$, then the system throughput function $\Lambda(V)$ is increasing concave with v_j, i.e. $\Lambda(V) + \Lambda(V + 2e_j) \leq 2\Lambda(V + e_j)$.
   Next, we change the service time distribution of infinite-server queues (IS) from exponential distribution to a general distribution without changing the mean. From insensitivity property in Lemma 3, $\Lambda(V)$ for the general distribution travel time case remains the same as that of the exponentially distributed travel time. Therefore, we conclude that $\Lambda(V) + \Lambda(V + 2e_j) \leq 2\Lambda(V + e_j)$ holds for closed queueing network in our setting.
2. As a result of Part 1 above, the first part of $h(V)$ is increasing concave with v_j, as $\bar{Z}$ is independent from V. Now, as $-\sum_{j\in F} c_j v_j$ is linear with v_j and the concavity is preserved under addition, the second part is concave with v_j. Moreover, $-\sum_{k\in S} \beta_k \alpha_k$ does not depend on V. Therefore, $h(V)$ is concave with v_j.
3. We now prove the second statement. From the first part of proof above, we know that $\Lambda(V)$ is increasing with v_j, $\forall j \in F$. Since $A_i(V) = \frac{\lambda_i}{\alpha_i}\Lambda(V)$, we conclude that $A_i(V)$ is increasing concave function in v_j for all $i, j \in F$.

Appendix D. Proof of Theorem 10

We first establish the supermodularity property (Topkis, 1998, p. 43) of the objective function below.

Lemma 13. *If $|F| = 2$, $V = (v_1, v_2)$, the objective function is supermodular, i.e., $h(v_1 + 1, v_2 - 1) + h(v_1, v_2) \leq h(v_1 + 1, v_2) + h(v_1, v_2 - 1)$.*

Proof. From Shanthikumar and Yao (1988a, Lemma 6 (ii)), the throughput is super modular in the case of exponential service time, i.e.

$$\Lambda(v_1 + 1, v_2 - 1) + \Lambda(v_1, v_2) \leq \Lambda(v_1 + 1, v_2) + \Lambda(v_1, v_2 - 1)$$

From Lemma 3, it also hold for any general travel time distribution. Multiplied it with the constant $\bar{Z}$, and minus $c_1(2v_1 + 1) + c_2(2v_2 - 1) + 2\sum_{k\in S} \beta_k \alpha_k$ on both sides of the inequality, we can conclude that the supermodular property also holds for function $h(v_1, v_2)$. □

We now prove Theorem 10 by transforming the problem and then using induction. Let $|V| = v_1 + v_2$. Consider the following transformed optimization problem:

$$\max_{V\in\mathbb{N}^2} h(V) \quad \text{such that } V \leq \hat{V}, |V| = n.$$

We exploit the supermodularity of h, proved in Lemma 13, and use induction on n to establish that for every n, Algorithm 2 outputs the optimal solution for the new constrained optimization problem above.

If $|V^*| = 3$, which means there is only one extra charger needed to be allocated on $V^1 = (1, 1)$. As the algorithm evaluates both $V = (2, 1)$ and $V = (1, 2)$ allocations and there is no other allocation, Algorithm 2 achieves the optimal solution.

Suppose the algorithm generates the optimal solution when $|V^*| = n$, denoted as $(a, n - a)$. For any $b < a$, we have

$$h(b, n - b) \leq h(a, n - a) \tag{D.1}$$

In order to prove the result, we only need to show that, when $|V^*| = n + 1$, the optimal solution is either $(a + 1, n - a)$ or $(a, n + 1 - a)$, which are evaluated in the algorithm. It suffices to show that any other allocation not evaluated by the algorithm $(b, n+1-b)$ cannot achieve a higher profit than either $(a+1, n-a)$ or $(a, n+1-a)$. We show the case with $b < a$, and the other case with $b > a + 1$ can be proved similarly.

As $h(v_1, v_2)$ is concave in v_2 from Theorem 9, we conclude

$$h(b, n+1-b) + h(b, n-a) \leq h(b, n-b) + h(b, n+1-a) \quad \text{(D.2)}$$

since $n - a < n + 1 - a \leq n - b < n + 1 - b$. From Lemma 13, $h(v_1, v_2)$ is supermodular. This implies

$$h(b, n+1-a) + h(a, n-a) \leq h(a, n+1-a) + h(b, n-a) \quad \text{(D.3)}$$

since $b < a$ and $n - a < n + 1 - a$. Adding up Eqs. (D.1), (D.2), and (D.3), we have

$$h(b, n+1-b) \leq h(a, n+1-a) \quad \text{(D.4)}$$

which concludes the result for case $b < a$ and finishes the induction.

Finally, we discuss the constraint $v_1 \leq \hat{v_1}$ and $v_2 \leq \hat{v_2}$ in (14). (i) If neither of the constraint is active, then the global maximum V^* found in Algorithm 2 is the optimal solution to the optimization problem. (ii) If only one constraint is activated, say $v_1 = \hat{v_1}$, then the optimization becomes univariate and Algorithm 2 generates the optimal solution, since the iteration continues as long as $h(v_1, v_2 + 1) > h(v_1, v_2)$. (iii) If both are active, then the upper bound $\hat{V}$ becomes the solution, from the last line in Algorithm 2.

References

Akyildiz, I. F., & Bolch, G. (1988). Mean value analysis approximation for multiple server queueing networks. *Performance Evaluation*, *8*(2), 77–91.

Ataç, S., Obrenović, N., & Bierlaire, M. (2021). Vehicle sharing systems: A review and a holistic management framework. *EURO Journal on Transportation and Logistics*, Article 100033.

Balsamo, S. (2000). Product form queueing networks. In *Performance evaluation: origins and directions* (pp. 377–401). Springer.

Balsamo, S., & Marin, A. (2007). Queueing networks. In *International school on formal methods for the design of computer, communication and software systems* (pp. 34–82). Springer.

Baskett, F., Chandy, K. M., Muntz, R. R., & Palacios, F. G. (1975). Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, *22*(2), 248–260.

Boyacı, B., Zografos, K. G., & Geroliminis, N. (2015). An optimization framework for the development of efficient one-way car-sharing systems. *European Journal of Operational Research*, *240*(3), 718–733.

Buzen, J. P. (1973). Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM*, *16*(9), 527–531.

Chandy, K. M., Howard, J. H., Jr., & Towsley, D. F. (1977). Product form and local balance in queueing networks. *Journal of the ACM*, *24*(2), 250–263.

Curry, G. L., & Feldman, R. M. (2010). *Manufacturing systems modeling and analysis*. Springer Science & Business Media.

Fanti, M. P., Mangini, A. M., Pedroncelli, G., & Ukovich, W. (2014). Fleet sizing for electric car sharing system via closed queueing networks. In *2014 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 1324–1329). IEEE.

Fox, B. (1966). Discrete optimization via marginal analysis. *Management Science*, *13*(3), 210–216.

Gelenbe, E., & Pujolle, G. (1998). *Introduction to queueing networks, Vol. 2*. Wiley New York.

George, D. K., & Xia, C. H. (2011). Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research*, *211*(1), 198–207.

Gupta, V., Harchol-Balter, M., Dai, J., & Zwart, B. (2010). On the inapproximability of M/G/K: why two moments of job size distribution are not enough. *Queueing Systems*, *64*(1), 5–48.

Hodgson, M. J. (1990). A flow-capturing location-allocation model. *Geographical Analysis*, *22*(3), 270–279.

Huang, Y., Li, S., & Qian, Z. S. (2015). Optimal deployment of alternative fueling stations on transportation networks considering deviation paths. *Networks and Spatial Economics*, *15*(1), 183–204.

Iglesias, R., Rossi, F., Zhang, R., & Pavone, M. (2019). A BCMP network approach to modeling and controlling autonomous mobility-on-demand systems. *International Journal of Robotics Research*, *38*(2–3), 357–374.

Illgen, S., & Höck, M. (2019). Literature review of the vehicle relocation problem in one-way car sharing networks. *Transportation Research, Part B (Methodological)*, *120*, 193–204.

Jung, J., Chow, J. Y., Jayakrishnan, R., & Park, J. Y. (2014). Stochastic dynamic itinerary interception refueling location problem with queue delay for electric taxi charging stations. *Transportation Research Part C (Emerging Technologies)*, *40*, 123–142.

Kaspi, M., Raviv, T., & Tzur, M. (2014). Parking reservation policies in one-way vehicle sharing systems. *Transportation Research, Part B (Methodological)*, *62*, 35–50.

Kumar, R. R., & Alok, K. (2020). Adoption of electric vehicle: A literature review and prospects for sustainability. *Journal of Cleaner Production*, *253*, Article 119911.

Lavenberg, S. (1983). *Computer performance modeling handbook*. Elsevier.

Lee, D., Quadrifoglio, L., Meloni, I., et al. (2016). Discovering relationships between factors of round-trip car sharing by using association rules approach. *Procedia Engineering*, *161*, 1282–1288.

Pavlenko, N., Slowik, P., & Lutsey, N. (2019). *When does electrifying shared mobility make economic sense?*: *Working paper*.

Raviv, T., & Kolka, O. (2013). Optimal inventory management of a bike-sharing station. *Iie Transactions*, *45*(10), 1077–1093.

Reiser, M., & Lavenberg, S. S. (1980). Mean-value analysis of closed multichain queuing networks. *Journal of the ACM*, *27*(2), 313–322.

Repoux, M., Kaspi, M., Boyacı, B., & Geroliminis, N. (2019). Dynamic prediction-based relocation policies in one-way station-based carsharing systems with complete journey reservations. *Transportation Research, Part B (Methodological)*, *130*, 82–104.

Serfozo, R. (2012). *Introduction to stochastic networks, Vol. 44*. Springer Science & Business Media.

Shanthikumar, J. G., & Yao, D. D. (1987). Optimal server allocation in a system of multi-server stations. *Management Science*, *33*(9), 1173–1180.

Shanthikumar, J. G., & Yao, D. D. (1988a). On server allocation in multiple center manufacturing systems. *Operations Research*, *36*(2), 333–342.

Shanthikumar, J. G., & Yao, D. D. (1988b). Second-order properties of the throughput of a closed queueing network. *Mathematics of Operations Research*, *13*(3), 524–534.

Shen, Z.-J. M., Feng, B., Mao, C., & Ran, L. (2019). Optimization models for electric vehicle service operations: A literature review. *Transportation Research, Part B (Methodological)*.

Smith, J. M. (2018). *Introduction to queueing networks: theory practice*. Springer.

Topkis, D. M. (1998). *Supermodularity and complementarity*. Princeton University Press.

Weikl, S., & Bogenberger, K. (2015). A practice-ready relocation model for free-floating carsharing systems with electric vehicles–mesoscopic approach and field trial results. *Transportation Research Part C (Emerging Technologies)*, *57*, 206–223.

Weldon, P., Morrissey, P., & O'Mahony, M. (2018). Long-term cost of ownership comparative analysis between electric vehicles and internal combustion engine vehicles. *Sustainable Cities and Society*, *39*, 578–591.

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operations Research*, *30*(2), 223–231.

Yang, J., Dong, J., & Hu, L. (2017). A data-driven optimization-based approach for siting and sizing of electric taxi charging stations. *Transportation Research Part C (Emerging Technologies)*, *77*, 462–477.