Queueing Systems
https://doi.org/10.1007/s11134-021-09707-w

Complete resource pooling of a load-balancing policy for a network of battery swapping stations

Fiona Sloothaak[1] · James Cruise[2] · Seva Shneer[3,4] · Maria Vlasiou[1,5] · Bert Zwart[1,6]

Received: 1 December 2020 / Revised: 16 April 2021 / Accepted: 26 April 2021

Abstract

To reduce carbon emission in the transportation sector, there is currently a steady move taking place to an electrified transportation system. This brings about various issues for which a promising solution involves the construction and operation of a battery swapping infrastructure rather than in-vehicle charging of batteries. In this paper, we study a closed Markovian queueing network that allows for spare batteries under a dynamic arrival policy. We propose a provisioning rule for the capacity levels and show that these lead to near-optimal resource utilization, while guaranteeing good quality-of-service levels for electric vehicle users. Key in the derivations is to prove a state-space collapse result, which in turn implies that performance levels are as good as if there would have been a single station with an aggregated number of resources, thus achieving complete resource pooling.

✉ Fiona Sloothaak
f.sloothaak@tue.nl

James Cruise
james.cruise@riverlane.com

Seva Shneer
v.shneer@hw.ac.uk

Maria Vlasiou
m.vlasiou@tue.nl

Bert Zwart
bert.zwart@cwi.nl

1 Eindhoven University of Technology, Eindhoven, The Netherlands

2 Riverlane, Cambridge, UK

3 Heriot-Watt University, Edinburgh, UK

4 Novosibirsk State University, Novosibirsk, Russia

5 University of Twente, Enschede, The Netherlands

6 Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

Published online: 01 June 2021

Keywords Stochastic model applications · Battery swapping · Markov Process · Energy

Mathematics Subject Classification 60K30 · 90B15 · 90B22

1 Introduction

A key challenge in the deployment and take-up of electric vehicles by society is the provision of a scalable charging infrastructure. A viable solution is the development of a battery swapping network. Currently, there has been work done on the operation and control of a single battery swapping station (for example [20]), but there is a clear gap within the literature when extending this to the operation of a wider network of stations. In this paper, we introduce a novel stochastic network model describing a network of battery swapping stations which clearly addresses this need and provides a foundation for future studies. In addition, we carry out a detailed analysis of this model and obtained a number of novel insights into the operation of a battery swapping network.

A steady energy transition is taking place due to the de-carbonization of the economy, leading to many intrinsic challenges and research opportunities, of which an overview is given in [2,14]. There are numerous challenging problems caused by developments on the demand side. Examples include control problems in local, smart distribution grids, as well as managing increasing demand irregularities caused by, for example, electric vehicles. Modeling the behavior of individual agents and their interaction naturally leads to stochastic models.

Despite the apparent need for alternative energy sources in the transportation sector, the adoption of electrified vehicles has been slow initially due to various practical challenges, such as high purchase costs of an EV, battery life problems and long battery charging times [17]. A possible solution to address these issues is the construction and operation of a battery swapping infrastructure. The upfront costs of purchase of an EV can be significantly reduced when battery swapping station operators own and lease batteries to customers, the batteries can be charged more appropriately to prolong batteries' lifetime [20], and EV users can experience a fast exchange of batteries in contrast to long charging times. Beyond the consumer benefits, the centralized charging paradigm of battery swapping allows the deferment of huge network reinforcement works required to support charging at home by connecting the chargers to the medium voltage network. Furthermore, the aggregation of a large number of batteries at charging stations can provide a comprehensive range of flexibility services to transmission and distribution network service operators.

In this paper, we introduce a model for EVs utilizing battery swapping technology within the context of a fixed region. Within the region there are a number of charging/swapping stations, and vehicles, in general, do not leave the region, leading to the conservation of batteries. This leads us to introduce a class of closed Markovian queueing network models, which we use in a novel way to model the evolution of the battery population within a city.

With the advancement of smartphones and online technologies, a range of service providers will utilize these advancements to provide occupancy level information to customers to improve delay performance. In a battery swapping system, such information can motivate EV users to visit the most appealing location in the direct vicinity. In this paper, we integrate a *load-balancing* policy to incorporate this. An intrinsic problem is to establish suitable capacity levels that account for the inherent tradeoff between EV users' quality-of-service and operational costs. To the best of our knowledge, this is the first work that considers this question for a battery swapping system in a network framework under a dynamic arrival policy.

Adequately balancing service performance and resource utilization is very much in the spirit of the *Quality-and-Efficiency-Driven (QED) regime* known from asymptotic many-server queueing theory [12]. Typically, this gives rise to a square-root slack provisioning policy for the capacity levels and has been successfully implemented in many applications such as call centers [4,13,26], healthcare systems [11,23,25] and more. This policy leads to favorable performance for large systems: as the number of customers r grows large, the waiting probability tends to a value strictly between zero and one, the waiting time vanishes with a rate $1/\sqrt{r}$, and near-optimal resource utilization of $1 - O(1/\sqrt{r})$ is achieved. To inherit such properties for the battery swapping framework, we adopt a similar capacity level design policy for both the number of charging servers and the number of spare batteries relative to the expected offered load under the load-balancing arrival strategy.

To add to the agreeable properties of delay performance in the QED regime, the arrival strategy ensures that the relative charging loads at the different stations do not grow apart too much since arriving EV users always move to the least loaded station. This phenomenon has been observed in a number of settings and is referred to as state-space collapse; see [5,24] for an overview and [9] for work most closely related to this paper. In fact, when capacity levels are chosen appropriately, this effect is so strong that complete resource pooling takes place: the system behaves as if there is only a single station with an aggregated number of resources. It ensures that it is unlikely that EV users are waiting for a battery at one station, while another is readily available at any other station, even among those stations that are far from his direct vicinity.

The first main contribution of this paper is the introduction of a stochastic model for battery charging in a network setting. In recent years, there has been a growing amount of research on both the planning/design as well as the operation/scheduling in battery swapping systems; see [20] for an overview. Most papers employ robust optimization techniques to find optimal solutions for certain objectives, while little of the works focus on the quality-of-service for EV users. The exception is a collection of papers written by a set of authors [16–20], that use asymptotic analysis and Markov Decision Process techniques to propose suitable solutions. Whereas the focus in those papers is on issues arising in a single station, we propose a network setting to account for queue length correlations between stations.

Our second main contribution involves the novelty of our load-balancing arrival mechanism. Load-balancing policies have attracted a lot of attention in recent years due to extremely relevant applications in large data centers; see [22] for an overview. Typically, these systems comprise many single-server stations where a central dispatcher decides where to allocate incoming tasks. In contrast, our framework involves

a network of (a fixed number of) multi-server stations for which we introduce a unique feature: an arriving EV user restricts itself to move only to one of the stations in his direct vicinity. By appropriately setting the capacity levels according to the QED provisioning rule, we show that this constraint becomes redundant in the sense that the resource pooling effect can still be achieved.

In this paper, we also make several theoretical contributions. Direct analysis of the steady-state distribution of the queue-length process is intractable under the load-balancing strategy in the case of multiple stations. Instead, we resort to a fluid and diffusion limit approach. We derive the existence of the fluid limit and point out its unique invariant state. Using a diffusion-scaled queue length process, we zoom in on the fluctuations around the invariant state. We prove a state-space collapse (SSC) result by showing that in the limit (as the number of EVs grows larger) the diffusion-scaled queue lengths tend to become arbitrarily close almost instantaneously and stay that way for any fixed interval. This property can be exploited to derive the limiting queue length behavior at every station, and show that it implies the complete resource pooling effect. The derivations of our results rely heavily on the framework developed by Dai and Tezcan [9], that in turn can be seen as an extension of [6]. We adapt their framework to incorporate a closed network setting under the novel load-balancing policy.

The introduction of the novel framework within this paper acts as a foundation for a substantial research program in the modeling of battery swapping networks. This will provide practitioners with a better understanding of how such networks should be designed and operated from both the perspective of quality of service requirements but also from an economic viewpoint. This can be carried out by enriching the model, here we highlight a few possible directions we consider important and challenging future steps. Each of these will provide a detailed insight into a specific aspect of such systems. Firstly, the inclusion of multiple customer types to model a range of car brands within the network using different battery systems. Secondly, there is a delay between the moment an EV user consults queue length information and the actual arrival due to transportation time. As is perceived in health care settings and bike-sharing systems, this can have a considerable effect on the queue length behavior. A third enhancement would be to incorporate a time-inhomogeneous demand rate to better simulate the expected diurnal variation. This will lead to a varying amount of slackness in the capacity within the QED regime. Finally, there is substantial underlying variability in the fluctuating energy prices, which sharply rise whenever the energy grid is more strained and vice versa. A battery swapping infrastructure will be sensitive to these prices changes and can provide an indispensable asset for supporting a stable grid in the future, since it can relieve strain during peak moments by deferring the moment of charging or even deplete batteries, providing energy to the grid. It is beyond the scope of this paper to provide efficient and adequate provisioning rules in these challenging settings, yet they offer intriguing avenues to pursue in future research. The main insight provided in the present study is the effectiveness of simple load-balancing policies, and while the model is parsimonious, this insight is useful in, at least, the planning stage of a swapping network.

The remainder of this paper is organized as follows: In Sect. 2, we describe the battery swapping network and its corresponding load-balancing arrival mechanism.

In Sect. 3, we present the fluid and diffusion results in the special case of a single station, and generalize these results for the multiple stations setting in Sect. 4. Our results imply approximations for certain performance measures, which we validate through several simulation experiments described in Sect. 5.

2 Model description

In this paper, we consider a queueing network with S battery swapping stations and r EVs. Each EV has one battery (collection) providing the energy for the car to drive. Every station $i \in \{1, \ldots, S\}$ has three types of assets: F_i *charging points*, B_i *spare batteries* and G_i *swapping servers*. Whenever there is an EV arrival at a station, a swapping server takes out the almost depleted battery and exchanges it for a fully charged one if available. The swapping time is relatively very short (with respect to charging times), and therefore, we assume it to occur instantaneously. Batteries in need of charging are being recharged whenever a charging point is available, and we assume every recharge to take an exponential amount of time with rate μ, independent of everything else. Whenever a battery is fully charged, it is placed in an EV immediately if one is waiting, and otherwise stocked for a future EV arrival. After receiving a fully charged battery, the EV requires recharging after an exponential amount of time with rate λ. With probability p_{ij} stations i and j are in the EV user's direct vicinity. We assume that EV users consult some online device, and are motivated to move to the station that is relatively least loaded (ties are broken evenly). We define which station is relatively least loaded more precisely later in this section. Figure 1 illustrates the closed queueing model under this load-balancing arrival mechanism.

We point out that batteries are always exchanged, and therefore, the number of batteries physically present at a station can never be below this station's number of spare batteries. In fact, this observation implies that the queueing model is closed, where the total number of batteries is given by

$$\text{Total \# batteries in system} = r + \sum_{j=1}^{S} B_i.$$

Another observation concerns the role of the swapping servers. Whenever a battery is taken out of the EV, it cannot move from the swapping server until an exchange of batteries has taken place. Thus, no more than $B_i + G_i$ batteries can be charged simultaneously at a station $i \in \{1, \ldots, S\}$. As a consequence, having more charging points creates no additional charging capacity, and can be bounded by

$$F_i \leq B_i + G_i, \quad i = 1, \ldots, S. \tag{1}$$

In addition, we assume that the number of such expensive swapping technologies is small at every station, i.e., $G_i < G$ for all $i = 1, \ldots, S$, with $G < \infty$ being a small fixed number.

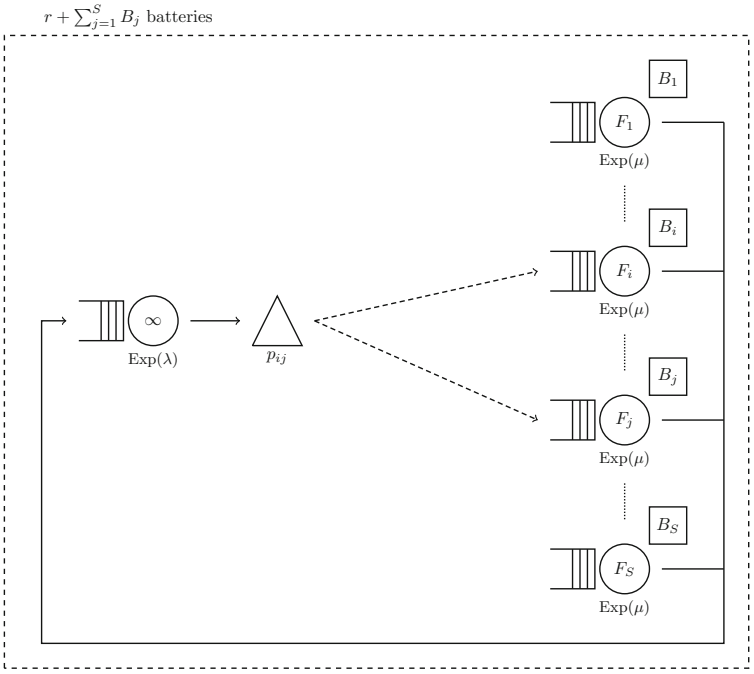

Fig. 1 Illustration of closed queueing network with multiple stations

The main quantity of interest in this paper is the number of batteries that are in need of charging, i.e., the aggregated number of batteries that are being charged at a charging point and the possible exchanged batteries that are waiting for an available charging point. We also refer to this quantity as the queue length. Let $Q_i(t)$ denote the number of batteries in need of charging at station i at time $t \geq 0$, and we write $Q(t) = (Q_1(t), \ldots, Q_S(t))$. Besides the queue length process, we focus on three performance measures in this paper: the waiting probability of an arbitrary EV, its expected waiting time and the resource utilization levels of the stations. As the role of swapping servers is non-existent in this framework, we consider the resources of the swapping stations to be the charging points and the spare batteries. We define the utilization level of the charging points to be the fraction of charging points that are busy with charging, and the utilization level of the spare batteries to be the fraction of batteries that are not fully charged with respect to the total number of batteries at the station. In steady state, the latter corresponds to the fraction of time at a station that a battery is expected to wait for an arriving EV.

To achieve favorable performance levels, we propose an associated QED-scaled capacity level for the resources at the stations. More specifically, we consider a sequence of systems indexed by the number of cars r, where we write a superscript r for processes and quantities to stress the dependency on r. Under the policy where every arrival would choose randomly between the two stations in its direct vicinity, we observe that $p_i = \sum_{j=1}^{S} p_{ij}/2$ represents the effective arrival probability for every

station $i = 1, \ldots, S$. Therefore, for a system with r cars, we set the capacity levels of the number of charging points and the number of spare batteries as

$$\begin{cases} B_i^r = p_i \left(\frac{\lambda r}{\mu} + \beta \sqrt{\frac{\lambda r}{\mu}} \right) & \beta \in \mathbb{R}, \\ F_i^r = p_i \left(\frac{\lambda r}{\mu} + \gamma \sqrt{\frac{\lambda r}{\mu}} \right) & \gamma \leq \beta, \end{cases} \tag{2}$$

for all $i = 1, \ldots, S$. We remark that the bound for the number of charging points originates from (1). Since the number of swapping servers is fixed and small and the number of cars r grows large, this condition reduces to the $\gamma \leq \beta$ requirement in (2).

Since there are two types of resources at every station, i.e., charging points and spare batteries, one can consider two types of utilization levels. However, in view of (2), we see that the capacity levels of both resources are of the magnitude $p_i \lambda r/\mu + O(\sqrt{r})$, and hence, the utilization levels of both resources are given by $Q(t)/(p_i \lambda r/\mu)(1 + o(1))$. Using this observation, we define the relative occupancy level (load) of a station as $Q_i(t)/p_i$. We let our load-balancing policy prescribe that an EV in need of charging closest to station i and j at time $t \geq 0$ moves to station i iff

$$\frac{Q_i(t)}{p_i} < \frac{Q_j(t)}{p_j}, \tag{3}$$

where ties are broken evenly. In our results, we show that this load-balancing policy ensures that the resource utilization levels at the different stations are approximately equal at all times. Consequently, this also ensures that the expected waiting times are approximately the same at every station at all times.

Remark 1 Our modeling prescribes that every EV user can choose between two stations in its direct vicinity. We point out that this is done for simplicity, as it helps to describe our scaling and load-balancing policy in a clear and concise manner. We point out that our model and results extends naturally to the cases where some EV arrivals may always move to one station, and some EV arrivals choose from multiple stations. With respect to the modeling, this extension can be included as follows: Let $\mathcal{M}$ be the set of arrival types, where every $m \in \mathcal{M}$ is a set of stations that isc in the direct vicinity of the EV user. Let s_m, $m \in \mathcal{M}$, denote the probability that an EV arrival is of type m. Then, the effective arrival rate at any station $i \in \{1, \ldots, S\}$ is given by

$$p_i = \sum_{m \in \mathcal{M}} \mathbb{1}_{\{i \in m\}} \, s_m / |m|.$$

In this extended setting, the scaling (2) for the number of resources and the load-balancing policy (3) remains the same.

3 System behavior in case of a single swapping station

When there is only a single battery swapping station, all EVs simply move to this station with probability one. The system reduces to a closed network where batteries

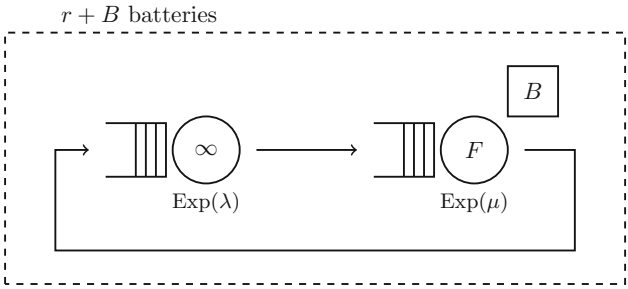

Fig. 2 Illustration of the closed queueing network with a single stations

are in two possible locations: either positioned in a car in no need of charging, or at the station. An illustration of the closed queueing model is given in Fig. 2. The square-root scaling rules reduces to

$$\begin{aligned} B^r &= \frac{\lambda r}{\mu} + \beta\sqrt{\frac{\lambda r}{\mu}}, \quad \beta \in \mathbb{R}, \\ F^r &= \frac{\lambda r}{\mu} + \gamma\sqrt{\frac{\lambda r}{\mu}}, \quad \gamma \leq \beta, \end{aligned} \tag{4}$$

where we suppress the subscript 1 for the station number in this case.

The notable advantage of a single station is that all resources are assembled at one entity, and inherently, no resources are unavailable by being at different locations. There is also a considerable upside in terms of the analysis: since there is no routing policy anymore, the queue length process becomes a simple birth–death process for which the steady-state distribution is easily derived. Yet, the steady-state distribution provides little qualitative insight into the queue length behavior, and in particular, the behavior of the process when it has not reached steady state yet. Therefore, we resort to fluid and diffusion limits, which in practice serve as good approximations for moderate to large-scale systems. This allows us to provide approximations for the performance measures of our interest, for example, the waiting probability and the expected waiting time.

At first glance, the single-station variant of our model may seem similar to the classic repair man model. This model and its QED-scaling implications are thoroughly treated in [10,11], which mainly focus on the healthcare setting. We point out that there is a crucial difference: our single-station model includes spare batteries, causing none of r cars to be waiting at the station as long as there are sufficient fully charged spares available. If $B = 0$, our model reduces to the repair man model with r machines and F repair men. Generally, however, the birth rates are different.

3.1 Steady-state distribution

As the queue length process is a birth–death process, it is straightforward to derive the steady-state distribution of the queue-length process by standard theory for Markov chains, irrespective of whether the QED scaled provisioning rules (4) hold. More specifically, the queue length $\{Q(t), t \geq 0\}$ is a birth–death process with state space $Q(t) \in \{0, 1, \ldots, B^r + r\}$ for all $t \geq 0$, with birth rate $\lambda(r - (Q(t) - B^r)^+)$ and death rate $\mu \min\{Q(t), F^r\}$. Let

$$\pi_k^{(B^r, F^r, r)} = \mathbb{P}(Q(\infty) = k)$$

denote the steady-state distribution of the number of batteries in need of charging.

Lemma 1 *Suppose $S = 1$, where the single swapping station has F charging points and B spare batteries, i.e., we disregard the scaling in (4). The steady-state distribution is given by*

$$\pi_k^{(B,F,r)} = \begin{cases} \frac{(\lambda r/\mu)^k}{k!}\pi_0^{(B,F,r)} & \text{if } 0 \leq k \leq \min\{B, F\}, \\ \frac{(\lambda r/\mu)^k}{F!F^{k-F}}\pi_0^{(B,F,r)} & \text{if } F \leq k \leq B, \\ \frac{r^B r!}{(r+B-k)!}\frac{(\lambda/\mu)^k}{k!}\pi_0^{(B,F,r)} & \text{if } B \leq k \leq F, \\ \frac{r^B r!}{(r+B-k)!}\frac{(\lambda/\mu)^k}{F!F^{k-F}}\pi_0^{(B,F,r)} & \text{if } \max\{B, F\} \leq k \leq B + r, \end{cases} \tag{5}$$

where, if $F \leq B$,

$$\pi_0^{(B,F,r)} = \left(\sum_{k=0}^{F} \frac{(\lambda r/\mu)^k}{k!} + \sum_{k=F+1}^{B-1} \frac{(\lambda r/\mu)^k}{F!F^{k-F}} + \sum_{k=B}^{B+r} \frac{r^B r!}{(r+B-k)!}\frac{(\lambda/\mu)^k}{F!F^{k-F}} \right)^{-1} \tag{6}$$

and, if $B \leq F$,

$$\pi_0^{(B,F,r)} = \left(\sum_{k=0}^{B} \frac{(\lambda r/\mu)^k}{k!} + \sum_{k=B+1}^{F-1} \frac{r^B r!}{(r+B-k)!}\frac{(\lambda/\mu)^k}{k!} + \sum_{k=F}^{B+r} \frac{r^B r!}{(r+B-k)!}\frac{(\lambda/\mu)^k}{F!F^{k-F}} \right)^{-1}. \tag{7}$$

Remark 2 In view of (1), we exclude the case that $F \geq B$ in our analysis further on in this paper. Yet, in an application where, for example, $G = \infty$ and hence $F \geq B$ possibly holds, we point out that the distribution can be derived similarly. That is, all EVs that arrive at the station find an available swapping server, and the swapping servers do not pose any restriction on the number of batteries that can be charged simultaneously. Only the number of charging points bounds the charging rate. One can also consider the QED provisioning rule in this case, which we treat in Appendix C of

the arXiv version of this paper [15]. Moreover, if both $G = \infty$ and $F = \infty$, Lemma 1 shows that

$$\pi_k^{(B,r)} = \begin{cases} \frac{(\lambda r/\mu)^k}{k!}\pi_0^{(B,r)} & \text{if } 0 \le k \le B, \\ \frac{r!r^B}{(r+B)!}\binom{r+B}{k}\left(\frac{\lambda}{\mu}\right)^k \pi_0^{(B,r)} & \text{if } B \le k \le B+r, \end{cases} \tag{8}$$

where

$$\pi_0^{(B,r)} = \left(\sum_{k=0}^{B-1} \frac{(\lambda r/\mu)^k}{k!} + \frac{r!r^B}{(r+B)!}\sum_{k=B}^{B+r}\binom{r+B}{k}\left(\frac{\lambda}{\mu}\right)^k\right)^{-1}. \tag{9}$$

Also in this particular case one can pose a QED provisioning rule for the number of spare batteries alone, and derive the asymptotic properties. We treat this case in Appendix B of the arXiv version of this paper [15].

3.2 Limiting queue length behavior

Due to the curse of dimensionality, it is very challenging to gain a qualitative insight in the (transient) behavior of processes in large-scale systems. Therefore, we resort to fluid and diffusion limits to provide good approximations for the behavior in the actual system when r is large. Recall that $Q^r(t)$ corresponds to the queue length process (the number of batteries in need of charging) under the scaling rules (4) with r cars at time $t \ge 0$. We consider the fluid scaling

$$\bar{Q}^r(t) = \frac{Q^r(t)}{r}, \quad r \ge 1, t \ge 0. \tag{10}$$

The fluid-scaled process converges to a deterministic, continuous monotone process with a single fixed steady-state value.

Proposition 1 *Suppose $S = 1$ and scaling rules (4) hold. If $\bar{Q}^r(0) \to \bar{Q}(0)$ as $r \to \infty$ with $\bar{Q}(0)$ a finite constant, then $\bar{Q}^r \to \bar{Q}$ in distribution as $r \to \infty$, where $\bar{Q}$ satisfies the ODE*

$$\frac{\mathrm{d}\bar{Q}(t)}{\mathrm{d}t} = \begin{cases} \lambda - \mu\bar{Q}(t) & \text{if } \bar{Q}(t) < \lambda/\mu, \\ \lambda^2/\mu - \lambda\bar{Q}(t) & \text{if } \bar{Q}(t) \ge \lambda/\mu, \end{cases}$$

and has the steady-state value

$$\lim_{t\to\infty} \bar{Q}(t) = \frac{\lambda}{\mu}.$$

Proposition 1 implies that the number of batteries in need of charging can be approximated by

$$Q^r(t) \approx r\,\bar{Q}(t),$$

where $\bar{Q}(t) = \lim_{r\to\infty} \bar{Q}^r(t)$ is a solution of an ODE. It describes the approximate (possible) transient behavior before reaching steady state. The proof of Proposition 1 is given in Appendix A of [15].

We point out that whenever the queue length is near its steady-state value, it remains close to its steady-state value from that time onward. That is, if $Q^r(t_0) \approx \lambda r/\mu$ for some $t_0 \geq 0$, then $Q^r(t) \approx \lambda r/\mu$ for all $t \geq t_0$. From that point on, the fluid limit becomes a rather rough estimate for the number of batteries in need of charging that allows for further investigation on the fluctuations around this value.

Therefore, we turn our focus to the diffusion scaling

$$\hat{Q}^r(t) = \frac{Q^r(t) - \lambda r/\mu}{\sqrt{\lambda r/\mu}}, \quad r \geq 1, t \geq 0. \tag{11}$$

This scaling provides more sensitive approximations, as it captures fluctuations of order $\sqrt{r}$. The diffusion-scaled process will tend to a piecewise linear Ornstein–Uhlenbeck processes, with a steady-state distribution that can be expressed analytically. The proof can be found in Appendix A of [15].

Theorem 1 *Suppose $S = 1$ and the system operates under (4). If $\hat{Q}^r(0) \to \hat{Q}(0)$ in distribution as $r \to \infty$, then $\hat{Q}^r \to \hat{Q}$ in distribution as $r \to \infty$. The process $\hat{Q}$ is a diffusion process with drift*

$$m(x) = -\lambda(x-\beta)^+ - \mu \min\{x, \gamma\},$$

and constant infinitesimal variance 2μ. The steady-state density of $\hat{Q}(\infty) = \lim_{t\to\infty} \hat{Q}(t)$ is given by

$$\hat{f}(x) = \begin{cases} \alpha_1 \frac{\phi(x)}{\Phi(\gamma)} & \text{if } x < \gamma, \\ \alpha_2 \left(\gamma e^{-\gamma(x-\gamma)}\right)\left(1 - e^{-\gamma(\beta-\gamma)}\right)^{-1} & \text{if } \gamma \leq x < \beta, \\ \alpha_3 \sqrt{\frac{\lambda}{\mu}} \phi\left(\frac{x-(\beta-\frac{\mu}{\lambda}\gamma)}{\sqrt{\mu/\lambda}}\right) \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1} & \text{if } x \geq \beta, \end{cases} \tag{12}$$

where $\alpha_i = r_i/(r_1 + r_2 + r_3)$, $i = 1, 2, 3$, with

$$r_1 = 1,$$
$$r_2 = \begin{cases} \phi(\gamma)\Phi(\gamma)^{-1} \frac{1}{\gamma}\left(1 - e^{-\gamma(\beta-\gamma)}\right) & \text{if } \gamma \neq 0, \\ \sqrt{\frac{2}{\pi}}\beta & \text{if } \gamma = 0, \end{cases}$$
$$r_3 = \frac{\phi(\gamma)}{\Phi(\gamma)} e^{-\gamma(\beta-\gamma)} \sqrt{\frac{\mu}{\lambda}} \phi\left(\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1} \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right).$$

Equation (12) in Theorem 1 is obtained by taking the limit of the scaled diffusion process (as $r \to \infty$), and finding its steady-state distribution (as $t \to \infty$). However, in order to obtain a good approximation of the steady-state distribution with a fixed number of cars r, it is arguably more reasonable to consider the steady-state distribution

of the scaled diffusion process (as $t \to \infty$) and next take the limit as $r \to \infty$. Fortunately, the following theorem shows that the order in which one takes the limits leads to the same result.

Theorem 2 *If $S = 1$ and (4) holds, the steady-state distribution of the diffusion scaled process $\hat{Q}^r(\infty)$ converges in distribution to $\hat{Q}(\infty)$ as in Theorem 1.*

The proof of Theorem 2 is given in Appendix A of [15]. As the order in which the limits are taken does not affect the result, we use the limiting process $\hat{Q}(\infty)$ to obtain approximations for the performance measures.

3.3 Performance measures

Typical performance measures for the QoS level for the EV users include the waiting probability and the expected waiting time. We view the efficiency-level for the station by the resources utilization. Typically, the QED regime in many-server systems causes the waiting probability to tend to a non-degenerate limit as $r \to \infty$, the waiting time to vanish, while the resource utilization tends to one. These features also appear in our system under the proposed QED scaling.

Due to the PASTA (Poisson Arrivals See Time Averages) property in open queueing systems where the arrival process is a time-homogeneous Poisson process, the steady-state value of any quantity is the same as at arrival instances. In particular, the waiting probability equals the steady-state probability that the number of fully charged batteries is zero, or equivalently, that the number of batteries in need of charging is at least B. Unfortunately, the arrival process in our closed setting is state-dependent. Yet, Theorem 1 shows that the fluctuations in arrival rate are of order $O(\sqrt{r})$, i.e., the arrival rate are $\lambda r - O(\sqrt{r})$ (with high probability). These small changes will therefore become negligible as $r \to \infty$. In other words, this argument implies that the PASTA property remains valid asymptotically. This notion can be formalized similarly as in [10]. Summarizing, if W denotes the waiting time of an arriving EV user, then

$$\mathbb{P}(W > 0) = \lim_{r\to\infty} \mathbb{P}\left(Q^r(\infty) \geq B^r\right) = \mathbb{P}\left(\hat{Q}(\infty) \geq \beta\right),$$

where $\hat{Q}(\infty)$ is as in Theorem 1.

The key concept to derive the expected waiting time is Little's law, stating that the long-term average number of waiting cars, denoted by Q_W^r, equals the long-term throughput multiplied by the average waiting time. In other words,

$$\mathbb{E}(Q_W^r) = \theta \mathbb{E}(W),$$

where the throughput θ can be viewed as the long-term average rate at which EVs arrive and hence also leave the battery swapping station. We can express the throughput as

$$\theta = \lambda r - \lambda \mathbb{E}(Q_W^r),$$

since the long-term average number of batteries not in need of charging is in fact the expected number of cars not waiting at the station in this closed system. Therefore, it follows that

$$\mathbb{E}(W) = \frac{\mathbb{E}(Q_W^r)}{\lambda(r - \mathbb{E}(Q_W^r))}. \tag{13}$$

In turn, the expected number of waiting cars can be derived directly using Theorem 1 and the observation that $Q_W^r = (Q^r(\infty) - B^r)^+$,

$$\begin{aligned}\mathbb{E}(Q_W^r) &= \sum_{k=B^r+1}^{r} (k - B^r)\mathbb{P}\left(Q(\infty) = k\right) \\ &= \sqrt{\frac{\lambda r}{\mu}} \sum_{k=B^r+1}^{r} \frac{k - B^r}{\sqrt{\lambda r/\mu}} \mathbb{P}\left(\hat{Q}(\infty) = \frac{k - \lambda r/\mu}{\sqrt{\lambda r/\mu}}\right) \\ &\sim \sqrt{\frac{\lambda r}{\mu}} \int_{\beta}^{\infty} (x - \beta)\hat{f}(x)\,\mathrm{d}x\end{aligned}$$

as $r \to \infty$. We point out that $\mathbb{E}(Q_W^r)$ is consequently of order $\Theta(\sqrt{r})$, and together with (13) this implies that $E(W)$ is of order $\Theta(1/\sqrt{r})$ and hence vanishes in the limit.

The resources will be fully utilized under (4) as $r \to \infty$. Theorem 1 implies that at most $O(\sqrt{r})$ charging points are not utilized, and the number of fully charged batteries is also of order $O(\sqrt{r})$. Therefore, as $r \to \infty$,

$$\rho_{F^r} = 1 - O(1/\sqrt{r}), \quad \rho_{B^r} = 1 - O(1/\sqrt{r}). \tag{14}$$

Theorem 3 *Suppose $S = 1$, and the system is operating under (4). Then the following properties hold as $r \to \infty$: The waiting probability has a non-degenerate limit given by*

$$\begin{aligned}\mathbb{P}(W > 0) \sim \mathbb{P}\left(\hat{Q}(\infty) \geq \beta\right) = \Bigg(1 + \sqrt{\frac{\lambda}{\mu}} \frac{\phi(\sqrt{\mu/\lambda}\gamma)}{\phi(\gamma)} e^{\gamma(\beta-\gamma)} \frac{\Phi(\gamma)}{\Phi(-\sqrt{\mu/\lambda}\gamma)} \\ + \sqrt{\frac{\lambda}{\mu}} \frac{\phi(\sqrt{\mu/\lambda}\gamma)}{\gamma} \left(e^{\gamma(\beta-\gamma)} - 1\right) \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1}\Bigg)^{-1}.\end{aligned}$$

The expected waiting time behaves as

$$\frac{\mathbb{E}(W)}{\sqrt{r}} \sim \frac{\alpha_3}{\sqrt{\lambda\mu}} \left(\sqrt{\frac{\mu}{\lambda}}\phi\left(\frac{\mu}{\lambda}\gamma\right) \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1} - \frac{\mu}{\lambda}\gamma\right),$$

with α_i are as in Theorem 1. Finally, the resource utilizations behave as

$$\rho_{F^r} \to 1, \quad \rho_{B^r} \to 1.$$

The proof of Theorem 3 is given in Appendix A of [15].

4 System behavior in case of multiple stations

When the number of stations $S \geq 2$, the analysis of system behavior needs to account for the underlying routing mechanism of arriving EVs. Whenever an EV is in need of recharging, stations i and j are in its direct vicinity with probability p_{ij}, and it chooses to move the station i if (3) holds. For a resource pooling effect to occur, we require that there is a sufficient number of pairs (i, j) for which $p_{ij} > 0$. For example, if the network consists of four stations with $p_{12} = p_{34} = 1/2$, there are no arrivals that can choose between one station in the set $\{1, 2\}$ and another in the set $\{3, 4\}$. Therefore, possible discrepancies in queue lengths are not leveled by the arrival mechanism between these two sets. Therefore, we assume that for every non-empty set $\mathcal{S}$ of stations, there is at least one pair (i, j) with $i \in \mathcal{S}$ and $j \notin \mathcal{S}$ for which $p_{ij} > 0$. This statement is equivalent to the following assumption.

Assumption 4 *Let $G = (V, E)$ be a graph, where $V = \{1, \ldots, S\}$ and $E = \{(i, j) : p_{ij} > 0\}$. We assume that the graph G is connected.*

Remark 3 For our results to follow through in the extended model as described in Remark 1, Assumption 4 needs to be updated as follows: Let $G = (V, E)$ be a graph, where $V = \{1, \ldots, S\}$ and $E = \{(i, j) : i, j \in m, |m| \geq 2, m \in \mathcal{M}\}$. Then, we assume that the graph G is connected. Note that if $m = 2$ for every $m \in \mathcal{M}$, the setting as well as this assumption reduces to the original setting as described in this paper.

4.1 System dynamics

There are many processes that are of interest in this system, and in particular, the queue length process at each station. In our analysis, we consider $\{\mathbb{X}^r(t), t \geq 0\}$ with

$$\mathbb{X}^r = \left(A^r, A^r_d, Q^r, Z^r, Y^r, T^r, D^r, L^r\right),$$

where

- $A^r = \left(A^r_{ij}; \{i, j\} \in E\right)$, where $A^r_{ij}(t)$ is the number of arrivals that are closest to stations i and j until time $t \geq 0$ in the rth system;
- $A^r_d = \left(A^r_{ij,i}; \{i, j\} \in E\right)$, where $A^r_{ij,i}(t)$ is the number of arrivals that are closest to stations i and j and are routed to station i until time $t \geq 0$ in the rth system;
- $Q^r = \left(Q^r_j; 1 \leq j \leq S\right)$, where $Q^r_j(t)$ is the number of batteries in need of charging at time $t \geq 0$ in the rth system;
- $Z^r = \left(Z^r_j; 1 \leq j \leq S\right)$, where $Z^r_j(t)$ is the number of busy servers (charging points) at time $t \geq 0$ in the rth system;
- Y^r, where $Y^r(t)$ is the aggregated time of all cars that are not waiting at some station until time $t \geq 0$ in the rth system;

- $T^r = \left(T_j^r; 1 \le j \le S\right)$, where $T_j^r(t)$ is the aggregated time of all servers at station j that were charging until time $t \ge 0$ in the rth system;
- $D^r = \left(D_j^r; 1 \le j \le S\right)$, where $D_j^r(t)$ is the number of service completions at station j until time $t \ge 0$ in the rth system;
- L^r, where $L^r(t)$ is the number of batteries that are positioned in an EV not waiting at a station in the rth system at time $t \ge 0$.

Clearly, there are strong relations between the individual processes in $\mathbb{X}^r$. For example, there is a routing policy that dictates where a car in need of a full battery drives to in order to swap its battery. This notion is captured by the arrival processes A^r (the classification of the different arrival types) and A_d^r (the routing decision). To generate the arrival and service completion processes, we introduce a set of independent Poisson processes. Let $\{\Lambda_{ij}(t), t \ge 0\}$ for all $\{i, j\} \in E$ be independent Poisson processes with rate $p_{ij}\lambda$ and $\{S_j(t), t \ge 1\}$ for all $1 \le j \le S$ be independent Poisson processes with rate μ. The system dynamics satisfy the following identities:

$$A_{ij}^r(t) = A_{ij,i}^r(t) + A_{ij,j}^r(t), \quad \forall \{i, j\} \in E, \tag{15}$$

$$A_{ij}^r(t) = \Lambda_{ij}\left(Y^r(t)\right), \quad \forall \{i, j\} \in E, \tag{16}$$

$$Q_j^r(t) = Q_j^r(0) + \sum_{i:\{i,j\}\in E} A_{ij,j}^r(t) - D_j^r(t), \quad \forall j = 1, \ldots, S, \tag{17}$$

$$D_j^r(t) = S_j\left(T_j^r(t)\right), \quad \forall j = 1, \ldots, S, \tag{18}$$

$$Y^r(t) = \int_0^t L^r(s)\,\mathrm{d}s, \tag{19}$$

$$T_j^r(t) = \int_0^t Z_j^r(s)\,\mathrm{d}s, \quad \forall j = 1, \ldots, S, \tag{20}$$

$$Z_j^r(t) = \min\{Q_j^r(t), F_j^r\}, \quad \forall j = 1, \ldots, S, \tag{21}$$

$$L^r(t) = r - \sum_{j=1}^{S}\left(Q_j^r(t) - B_j^r\right)^+, \tag{22}$$

$$\forall \{i, j\} \in E,\ A_{ij,i}^r(t) \text{ can only increase when } Q_i^r(t)/p_i \le Q_j^r(t)/p_j. \tag{23}$$

We refer to these equations as the system identities, and they prove to be central for deriving our results. The derivations use the framework set out in [9], which in turn is based on [6]. We adopt much of the notation and definitions in this paper, and before stating our main results, we repeat them for the purpose of self-containment. For each positive integer d, we denote by $\mathbb{D}^d[0, \infty]$ the d-dimensional Skorohod path space. For $x, y \in \mathbb{D}^d[0, \infty]$ and $T > 0$, let

$$\|x(\cdot) - y(\cdot)\|_T = \sup_{0\le t\le T} |x(t) - y(t)|,$$

where $|z| = \max_{i=1,\ldots,d} |z_i|$ for any $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$. The space $\mathbb{D}^d[0, \infty]$ is endowed with the J_1 topology, and the weak convergence in this space is considered with respect to this topology. We say a sequence of functions $\{x_n\} \in \mathbb{D}^d[0, \infty]$ converges uniformly on compact sets (u.o.c) sets to $x \in \mathbb{D}^d[0, \infty]$ as $n \to \infty$ if, for each $T \geq 0$,

$$\|x_n(\cdot) - x(\cdot)\|_T \to 0$$

as $n \to \infty$. Moreover, we say that $t \geq 0$ is a regular point of a function x if x is differentiable at $t \geq 0$, and denote its derivative by $x'(\cdot)$. We assume that the random variables in $\mathbb{X}^r$ live on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Often, we consider sample paths of stochastic processes, and whenever we want to make the dependence on the sample path explicit, we write $X^r(\cdot, \omega)$ for the sample path associated with $\omega \in \Omega$ for a stochastic process X^r.

4.2 Fluid limit

To capture the rough system dynamics, we consider the fluid-scaled process

$$\bar{\mathbb{X}} = \lim_{r \to \infty} \bar{\mathbb{X}}^r, \quad \bar{\mathbb{X}}^r = \frac{\mathbb{X}^r}{r}.$$

For each process X^r in $\mathbb{X}^r$, we define similarly its fluid equivalent as $\bar{X}^r = X^r/r$ and its limiting process $\bar{X} = \lim_{r \to \infty} \bar{X}^r$. We adopt the definition of a fluid limit and its invariant state(s) from [9]. That is, we consider $\mathcal{A} \subset \Omega$ such that the FSLLN holds, i.e.,

$$\frac{A_{ij}(rx)}{r} \to p_{ij}\lambda x, \quad \{i, j\} \in E \quad \text{and} \quad \frac{S_j(rx)}{r} \to \mu x, \quad j = 1, \ldots, S,$$

u.o.c. as $r \to \infty$. Due to the FSLLN, we observe that one can choose $\mathcal{A}$ large enough such that $\mathbb{P}(\mathcal{A}) = 1$.

Definition 1 We call $\bar{\mathbb{X}}$ a fluid limit of $\{\mathbb{X}^r\}$ if there exists an $\omega \in \mathcal{A}$ and (sub)sequence $\{r_l\}$ with $r_l \to \infty$ as $l \to \infty$, such that $\bar{\mathbb{X}}^{r_l}(\cdot, \omega)$ converges u.o.c. to $\bar{\mathbb{X}}(\cdot, \omega)$. Moreover, let $q = (q_1, \ldots, q_S)$ be an invariant state of the fluid limits if for any fluid limit $\bar{\mathbb{X}}$, $\bar{Q}(0) = (\bar{Q}_1(0), \ldots, \bar{Q}_S(0)) = (q_1, \ldots, q_S) = q$ implies that $\bar{Q}(t) = q$ for all $t \geq 0$.

In Proposition 1, we focus on the fluid-scaled queue length process only for $S = 1$, and the sequence $r_l = l$. Instead of requiring $\bar{Q}^r(0) \to \bar{Q}(0)$ with $\bar{Q}(0)$ a finite constant, Definition 1 allows for $\bar{Q}(0)$ to be random. Proposition 1 implies that in case that $S = 1$, the fluid limits exist and are deterministic, (Lipschitz) continuous paths that depend only on the realization of $\bar{Q}(0)$. Moreover, there is a single unique invariant state given by λ/μ. A similar result holds when $S \geq 2$.

Theorem 5 *Let $\{\mathbb{X}^r\}$ be a sequence of systems. Then the fluid limits exist, where each component is Lipschitz continuous. Each fluid limit $\bar{\mathbb{X}}$ satisfies the following equations for all $t \geq 0$:*

$$\bar{A}_{ij}(t) = \bar{A}_{ij,i}(t) + \bar{A}_{ij,j}(t), \quad \forall \{i, j\} \in E, \tag{24}$$

$$\bar{A}_{ij}(t) = p_{ij}\lambda \bar{Y}(t), \quad \forall \{i, j\} \in E, \tag{25}$$

$$\bar{Q}_j(t) = \bar{Q}_j(0) + \sum_{i:\{i,j\}\in E} \bar{A}_{ij,j}(t) - \bar{D}_j(t), \quad \forall j = 1, \ldots, S, \tag{26}$$

$$\bar{D}_j(t) = \mu \bar{T}_j(t), \quad \forall j = 1, \ldots, S, \tag{27}$$

$$\bar{Y}(t) = \int_0^t \bar{L}(s)\,\mathrm{d}s, \quad \forall j = 1, \ldots, S, \tag{28}$$

$$\bar{T}_j(t) = \int_0^t \bar{Z}_j(s)\,\mathrm{d}s, \quad \forall j = 1, \ldots, S, \tag{29}$$

$$\bar{Z}_j(t) = \min\{\bar{Q}_j(t), p_j\lambda/\mu\}, \quad \forall j = 1, \ldots, S, \tag{30}$$

$$\bar{L}(t) = 1 - \sum_{j=1}^{S} \left(\bar{Q}_j(t) - p_j\lambda/\mu\right)^+. \tag{31}$$

Also, for every $\{i, j\} \in E$, if t is a regular point of $\bar{\mathbb{X}}$, then

$$\bar{A}'_{ij,i}(t) = \lambda p_{ij} \bar{L}(t) \quad \textit{and} \quad \bar{A}'_{ij,j}(t) = 0 \quad \textit{if} \quad \frac{\bar{Q}_j(t)}{p_j} > \frac{\bar{Q}_i(t)}{p_i}. \tag{32}$$

Finally, there is a unique invariant state given by $q = (q_1, \ldots, q_S)$ with $q_i = p_i\lambda/\mu$ for $i = 1, \ldots, S$.

The (uniqueness of the) invariant state result for the fluid limit is central for the existence of a properly defined diffusion process as it states that if $\bar{Q}(0) = (p_1\lambda/\mu, \ldots, p_S\lambda/\mu)$, the fluid limits are time invariant. We present a proof of Theorem 5 in Appendix A.

4.3 Diffusion limit

Due to the policy governing which station a car drives to in order to replace a battery, one observes the so-called load-balancing effect. By setting the number of resources as in (2), this load-balancing effect is so strong that in fact complete resource pooling occurs. In other words, the system behaves as if there is a single large swapping station where the number of resources equals the aggregated total of the individual stations. This appealing consequence ensures that there are no idle resources at one station, while at another there are possible long waiting lines of cars that are waiting for a battery exchange.

The key concept to derive this effect is to show a state-space collapse (SSC) result. That is, we consider the diffusion-scaled queue length process defined as

$$\hat{Q}_i^r(t) = \frac{Q_i^r(t) - p_i \lambda r/\mu}{p_i \sqrt{\lambda r/\mu}}, \quad i = 1, \ldots, S.$$

In our model, the SSC result states that (almost instantaneously) the diffusion-scaled queue length processes are arbitrarily close at all stations, and stay close during any fixed interval.

Theorem 6 *Suppose*

$$\hat{Q}^r(0) \xrightarrow{d} \hat{Q}(0),$$

as $r \to \infty$, *where* $\hat{Q}(0)$ *is a random vector. Then, for every* $K^r = o(\sqrt{r})$ *with* $K^r \to \infty$ *as* $r \to \infty$, *and for every* $T > 0$ *and* $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{K^r/\sqrt{r} \le t \le T} |\hat{Q}_i(t) - \hat{Q}_j(t)| > \epsilon \right) \to 0 \tag{33}$$

for every $i, j \in \{1, \ldots, S\}$ *as* $r \to \infty$. *If, in addition, for every* $i, j \in \{1, \ldots, S\}$,

$$|\hat{Q}_i^r(0) - \hat{Q}_j^r(0)| \xrightarrow{\mathbb{P}} 0,$$

then

$$\mathbb{P}\left(\|\hat{Q}_i(\cdot) - \hat{Q}_j(\cdot)\|_T > \epsilon \right) \to 0 \tag{34}$$

for every $i, j \in \{1, \ldots, S\}$ *as* $r \to \infty$.

The proof of Theorem 6 is given in Appendix B. This result reveals that instead of considering the individual queue length processes, it suffices to track the total queue length process instead. More specifically, define the sequence of random processes $\{Q_\Sigma^r(t), t \ge 0\}$ with $r \in \mathbb{N}$, where $Q_\Sigma^r(t) = \sum_{j=1}^S Q_j^r(t)$, and

$$\hat{Q}_\Sigma^r(t) = \frac{\sum_{j=1}^S (Q_j^r(t) - p_j \lambda r/\mu)}{\sqrt{\lambda r/\mu}} = \sum_{j=1}^S p_j \hat{Q}_j^r(t).$$

As the state-space collapse implies that $\hat{Q}_i^r(t) \approx \hat{Q}_j^r(t)$ for all $i, j \in \{1, \ldots, S\}$ (for $t \ge K^r/\sqrt{r}$), we can approximate the queue length at an individual queue by

$$Q_j^r(t) = p_j \left(\frac{\lambda r}{\mu} + \hat{Q}_j^r(t) \sqrt{\frac{\lambda r}{\mu}} \right) \approx p_j \left(\frac{\lambda r}{\mu} + \hat{Q}_\Sigma^r(t) \sqrt{\frac{\lambda r}{\mu}} \right)$$

for all $j = 1, \ldots, S$. The limiting process of the total queue length can be derived using the SSC result.

Theorem 7 *Suppose* $\hat{Q}^r(0) \to \hat{Q}(0)$ *in distribution as* $r \to \infty$, *and*

$$\left| \hat{Q}_i^r(0) - \hat{Q}_j^r(0) \right| \xrightarrow{\mathbb{P}} 0$$

for all $i, j \in \{1, \ldots, S\}$. *Then,* $\hat{Q}_\Sigma^r \to \hat{Q}_\Sigma$ *in distribution as* $r \to \infty$, *where* $\hat{Q}_\Sigma$ *is a diffusion process with drift*

$$m(x) = -\lambda(x - \beta)^+ - \mu \min\{x, \gamma\},$$

and constant infinitesimal variance 2μ. *The steady-state density* $\hat{Q}_\Sigma(\infty)$ *is given by*

$$\hat{f}_\Sigma(x) = \begin{cases} \alpha_1 \frac{\phi(x)}{\Phi(\gamma)} & \text{if } x < \gamma, \\ \alpha_2 \left(\gamma e^{-\gamma(x-\gamma)}\right) \left(1 - e^{-\gamma(\beta-\gamma)}\right)^{-1} & \text{if } \gamma \le x < \beta, \\ \alpha_3 \sqrt{\frac{\lambda}{\mu}} \phi\left(\frac{x-(\beta-\frac{\mu}{\lambda}\gamma)}{\sqrt{\mu/\lambda}}\right) \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1} & \text{if } x \ge \beta, \end{cases} \tag{35}$$

where $\alpha_i = r_i/(r_1 + r_2 + r_3)$, $i = 1, 2, 3$, *with*

$$\begin{aligned} r_1 &= 1, \\ r_2 &= \begin{cases} \phi(\gamma)\Phi(\gamma)\frac{1}{\gamma}\left(1 - e^{-\gamma(\beta-\gamma)}\right) & \text{if } \gamma \neq 0, \\ \sqrt{\frac{2}{\pi}}\beta & \text{if } \gamma = 0, \end{cases} \\ r_3 &= \frac{\phi(\gamma)}{\Phi(\gamma)} e^{-\gamma(\beta-\gamma)} \sqrt{\frac{\mu}{\lambda}} \phi\left(\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1} \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right). \end{aligned}$$

Proof We observe that the steady-state density is a direct consequence of the diffusion process [7]. What remains to be shown is that $\hat{Q}_\Sigma^r$ converges to the described diffusion process as $r \to \infty$. Equivalently, we need to show that

$$d\hat{Q}_\Sigma(t) = -\lambda \left(\hat{Q}_\Sigma(t) - \beta\right)^+ - \mu \min\left\{\hat{Q}_\Sigma(t), \gamma\right\} + \sqrt{2\mu} dW(t), \tag{36}$$

where $\{W(t), t \ge 0\}$ is a standard Brownian motion. We note that, due to the system identities,

$$\hat{Q}_\Sigma(t) = \hat{Q}_\Sigma(0) + \frac{\sum_{\{i,j\} \in E} A_{ij}^r(t) - \sum_{j=1}^S D_j^r(t)}{\sqrt{\lambda r/\mu}},$$

where

$$\sum_{\{i,j\} \in E} A_{ij}^r(t) = \Lambda\left(\int_0^t L^r(s)\, \mathrm{d}s\right),$$

and

$$\sum_{j=1}^{S} D_j^r(t) = \sum_{j=1}^{S} S_j \left(\int_0^t Z_j^r(s)\,\mathrm{d}s \right).$$

We observe that, due to the FCLT and Theorem 5,

$$\frac{A\left(\int_0^t L^r(s)\,\mathrm{d}s\right) - \lambda \int_0^t L^r(s)\,\mathrm{d}s}{\sqrt{\lambda r/\mu}} = \frac{A\left(r\int_0^t \bar{L}^r(s)\,\mathrm{d}s\right) - \lambda r \int_0^t \bar{L}^r(s)\,\mathrm{d}s}{\sqrt{1/\mu}\sqrt{\lambda r}} \xrightarrow{d} \mathrm{BM}_A(t),$$

where $\{\mathrm{BM}_A(t), t \geq 0\}$ is Brownian motion with mean zero and variance μ. Similarly, due to the FCLT and Theorem 5,

$$\sum_{j=1}^{S} \frac{S_j\left(\int_0^t Z_j^r(s)\,\mathrm{d}s\right) - \mu \int_0^t Z_j^r(s)\mathrm{d}s}{\sqrt{\lambda r/\mu}} = \sum_{j=1}^{S} \frac{S_j\left(r\int_0^t \bar{Z}^r(s)\,\mathrm{d}s\right) - \mu r \int_0^t \bar{Z}^r(s)\mathrm{d}s}{\sqrt{1/(p_j\mu)}\sqrt{p_j\lambda r}} \xrightarrow{d} \mathrm{BM}_D(t),$$

where $\{\mathrm{BM}_D(t), t \geq 0\}$ is an (independent) Brownian motion with mean zero and variance μ. The sum of these two processes is equal (in distribution) to a Brownian with mean zero and variance 2μ, which contributes to the $\sqrt{2\mu}\,dW(t)$ term in (36). Next, we observe that, due to the system identities and the definition of the diffusion scaling,

$$\frac{\lambda \int_0^t L^r(s)\,\mathrm{d}s - \sum_{j=1}^{S} \mu \int_0^t Z_j^r(s)\,\mathrm{d}s}{\sqrt{\lambda r/\mu}} = -\lambda \sum_{j=1}^{S} p_j \int_0^t \left(\hat{Q}_j^r(s) - \beta\right)^+ \mathrm{d}s - \mu \sum_{j=1}^{S} p_j \int_0^t \min\left\{\hat{Q}_j^r(s), \gamma\right\} \mathrm{d}s.$$

Since

$$\min_{1\leq j\leq S} \hat{Q}^r(t) \leq \hat{Q}_\Sigma^r(t) \leq \max_{1\leq j\leq S} \hat{Q}^r(t)$$

for all $t \in [0, T]$, and due to Theorem 6,

$$\left\| \hat{Q}_\Sigma^r(\cdot) - \hat{Q}_j^r(\cdot) \right\|_T \leq \epsilon(r), \quad j = 1, \ldots, S.$$

where the sequence $\{\epsilon(r), r \in \mathbb{N}\}$ can be chosen such that $\epsilon(r) \to 0$ as $r \to \infty$. Then, as $r \to \infty$,

$$\frac{\lambda \int_0^t L^r(s)\,\mathrm{d}s - \sum_{j=1}^S \mu \int_0^t Z_j^r(s)\,\mathrm{d}s}{\sqrt{\lambda r/\mu}} \xrightarrow{\mathbb{P}} -\lambda \int_0^t \left(\hat{Q}_\Sigma(s) - \beta\right)^+ \mathrm{d}s - \mu \int_0^t \min\left\{\hat{Q}_\Sigma(s), \gamma\right\} \mathrm{d}s.$$

This contributes to the first two terms in (36). Applying the continuous mapping theorem concludes the proof. □

Another consequence of the state-space collapse result is that the waiting probabilities and expected waiting times are equal at all stations, as well as the resource utilization levels. In fact, it exhibits the same behavior as if there were a single station due to the complete resource pooling effect.

Corollary 1 *Suppose the system is operating under* (4)*. Then the following properties hold as $r \to \infty$ for all $i = 1, \ldots, S$: The waiting probability has a non-degenerate limit given by*

$$\mathbb{P}(W_i^r > 0) \to \mathbb{P}\left(\bar{Q}_\Sigma(\infty) \geq \beta\right) = \left(1 + \sqrt{\frac{\lambda}{\mu}} \frac{\phi(\sqrt{\mu/\lambda}\gamma)}{\phi(\gamma)} e^{\gamma(\beta-\gamma)} \frac{\Phi(\gamma)}{\Phi(-\sqrt{\mu/\lambda}\gamma)} + \sqrt{\frac{\lambda}{\mu}} \frac{\phi(\sqrt{\mu/\lambda}\gamma)}{\gamma} \left(e^{\gamma(\beta-\gamma)} - 1\right) \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1}\right)^{-1}.$$

The expected waiting time behaves as

$$\frac{\mathbb{E}(W_i^r)}{\sqrt{r}} \to \frac{\alpha_3}{\sqrt{\lambda\mu}} \left(\sqrt{\frac{\mu}{\lambda}} \phi\left(\frac{\mu}{\lambda}\gamma\right) \Phi\left(-\sqrt{\frac{\mu}{\lambda}}\gamma\right)^{-1} - \frac{\mu}{\lambda}\gamma\right),$$

with α_i as in Theorem 7. Finally, the resource utilizations behave as

$$\rho_{F_i^r} \to 1, \quad \rho_{B_i^r} \to 1.$$

5 Simulation experiments

The results presented are given in an asymptotic regime where the charging times are exponentially distributed. In this section, we conduct simulation experiments to evaluate the quality of our approximations and the robustness of the state-space collapse result. We first focus on a large-scale system to illustrate the implications of our results. Next, we also zoom in on a moderate-sized system that reflects a more realistic setting for an EV battery swapping infrastructure.

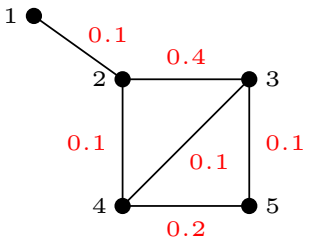

Fig. 3 Illustration of the battery swapping network with arrival streams used in the simulation experiments

5.1 Large-scale system

Throughout the experiments in this section, we consider a network with five stations where the arrival probabilities are given by $p_{12} = p_{24} = p_{34} = p_{35} = 0.1$, $p_{23} = 0.4$ and $p_{45} = 0.2$; see Fig. 3. This results in an effective arrival probability $p = (p_1, \ldots, p_5) = (0.05, 0.3, 0.3, 0.2, 0.15)$ at the stations.

As a battery swapping infrastructure currently does not exist yet in real-life, there is no (significant) data that can be exploited to obtain useful parameter choices. Instead, we discuss an adequate provisioning strategy under the following assumptions: We assume that the battery swapping facility installed (relatively) fast charging points where recharging takes 1 h on average ($\mu = 1$), and that every EV user returns for recharging services after every 40 h on average ($\lambda = 0.025$). In addition, we stress that our results are based on an asymptotic regime, and therefore require the system to be sufficiently large for the approximation to become meaningful. We allow for (at least) $r = 50{,}000$ EV users in this infrastructure. The effective loads at the stations in this case are

$$\left(\frac{p_j \lambda r}{\mu}\right)_{j=1,\ldots,5} = (62.5, 375, 375, 250, 187.5),$$

and we note that due to the QED provisioning rule in (2), the numbers of charging points and spare batteries are close to these values. Obviously, the number of resources are integer values, and in our simulation experiments we choose

$$F^r = \left(\left\lceil \frac{\lambda r}{\mu} + \gamma\sqrt{\frac{\lambda r}{\mu}} \right\rceil\right)_{j=1,\ldots,5}, \qquad B^r = \left(\left\lceil \frac{\lambda r}{\mu} + \beta\sqrt{\frac{\lambda r}{\mu}} \right\rceil\right)_{j=1,\ldots,5}.$$

5.1.1 State-space collapse for exponential charging times

A first-order approximation for the queue length process is implied by the fluid result in Theorem 5. We validate this approximation for the above-described setting, with initial queue length $Q_i^r(0) = 150$ for all stations $i = 1, \ldots, 5$. That is, only station 1 is initially overloaded, while all other station are underloaded. The equations in Theorem 5 together with the Lipschitz continuity describe a unique fluid limit with the given initial queue length. This yields the approximations

$$Q_i^r(t) \approx r\,\bar{Q}_(t), \quad j = 1, \ldots, S.$$

In particular, in the case when the initial queue length is $Q_i^r(0) = 150$ for all stations $i = 1, \ldots, 5$, this results in the approximations

$$Q_1^r(t) \approx \begin{cases} 150 - 62.5t & \text{if } t \le t_3, \\ 62.5e^{1.4-t} & \text{if } t_3 \le t \le t_4, \\ 62.5 + \frac{195}{64}e^{1.4-t} - \frac{517}{16}e^{-t} & \text{otherwise,} \end{cases}$$

$$Q_2^r(t) \approx Q_3^r(t) \approx \begin{cases} 498.5 + 0.625t - 348.5e^{-t} & \text{if } t \le t_2, \\ \frac{75t}{152} + \frac{14955}{38} - \frac{7755}{38}e^{-t} & \text{if } t_2 \le t \le t_3, \\ \frac{7500}{19} - \frac{75}{152}e^{1.4-t} - \frac{7755}{38}e^{-t} & \text{if } t_3 \le t \le t_4, \\ 375 + \frac{585}{32}e^{1.4-t} - \frac{1551}{8}e^{-t} & \text{otherwise,} \end{cases}$$

$$Q_4^r(t) \approx \begin{cases} 249.25 + 0.3125t - 99.25e^{-t} & \text{if } t \le t_1, \\ \frac{997}{7} + \frac{5}{28}t + 29e^{-t} & \text{if } t_1 \le t \le t_2, \\ \frac{4985}{19} + \frac{25}{76}t - \frac{2585}{19}e^{-t} & \text{if } t_2 \le t \le t_3, \\ \frac{5000}{19} - \frac{25}{76}e^{1.4-t} - \frac{2585}{19}e^{-t} & \text{if } t_3 \le t \le t_4, \\ 250 + \frac{195}{16}e^{1.4-t} - 129.25e^{-t} & \text{otherwise,} \end{cases}$$

$$Q_5^r(t) \approx \begin{cases} 150e^{-t} & \text{if } t \le t_1, \\ \frac{15t}{112} + \frac{2991}{28} + \frac{87}{4}e^{-t} & \text{if } t_1 \le t \le t_2, \\ \frac{14955}{76} + \frac{75}{304}t - \frac{7755}{76}e^{-t} & \text{if } t_2 \le t \le t_3, \\ \frac{3750}{19} - \frac{75}{304}e^{1.4-t} - \frac{7755}{76}e^{-t} & \text{if } t_3 \le t \le t_4, \\ 187.5 + \frac{585}{64}e^{1.4-t} - \frac{1551}{16}e^{-t} & \text{otherwise,} \end{cases}$$

where $t_1 \approx 0.1826$, $t_2 \approx 0.3189$, $t_3 = 1.4$ and $t_4 \approx 1.4758$. The times t_i, $i = 1, 2, 4$, correspond to the times where two stations (approximately) have the same relative queue lengths, and t_3 is the moment where the number of EVs in need of recharging is (approximately) equal to the number of stations/spare batteries.

A sample path comparison with its fluid approximation (dotted lines) is graphically illustrated in Fig. 4. We observe that the fluid limit approximations capture the typical values of the actual queue length process quite accurately. We observe apparent fluctuations around its approximation, and we note that these become relatively small as r grows large.

To observe the state-space collapse, we plot the same sample path in its diffusion scaling; see Fig. 5. Indeed, around $t_4 \approx 1.4758$ the diffusion-scaled queue lengths appear to become close and remain nearly equal to one another after this time. In addition, as time moves, the diffusion-scaled queue lengths fluctuate around zero.

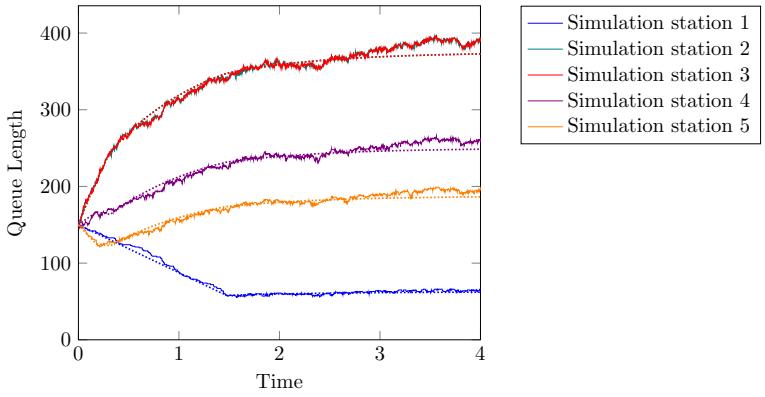

Fig. 4 Sample path of the queue lengths when $Q^r(0) = (150, 150, 150, 150, 150)$

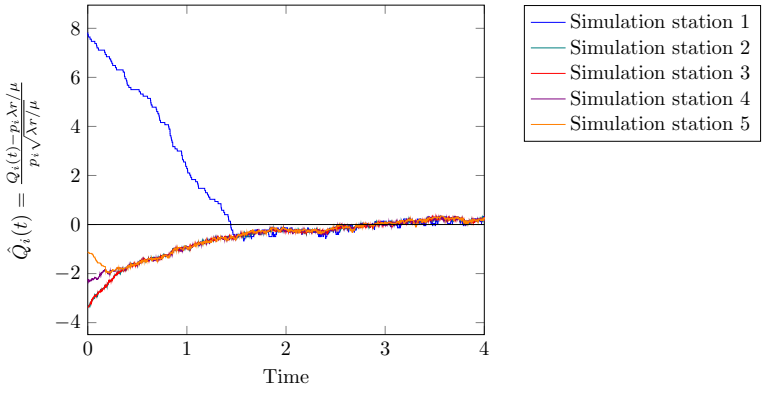

Fig. 5 Sample path of the diffusion-scaled queue lengths with starting point $Q^r(0) = (150, 150, 150, 150, 150)$

5.1.2 Performance measures

Our results imply approximations for performance measures such as the waiting probability and waiting time; see Corollary 1. In particular, the state-space collapse result implies that the performance at all stations is approximately the same, and can be approximated by the closed-form expressions as given in Corollary 1.

In Fig. 6, we plotted the waiting probabilities of all stations in the case of 2,500,000 EV arrivals averaged over 20 samples for the large-scaled system. We point out that the stair-type effect appearing in the waiting probabilities is due to the ceiling of the number of resources at the stations. Moreover, as r is finite and we use the ceiling function, the waiting times are not all exactly equal, which is most apparent for station 1. This is also reflected in Fig. 5, where a closer view suggests that the diffusion-scaled queue length at station 1 is smaller than the queue length at another station (often station 2 or station 3). As r grows large, the waiting probabilities do grow closer and move near to their asymptotic expressions. Still, the waiting probabilities are typically below their

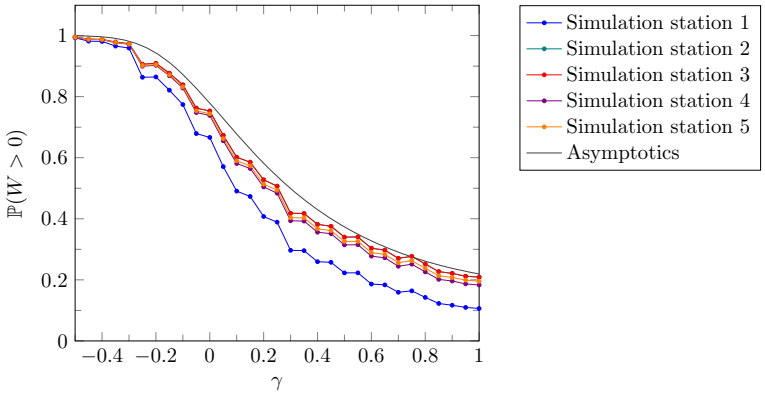

Fig. 6 Waiting probabilities with respect to its asymptotic expression when $\beta = 1$

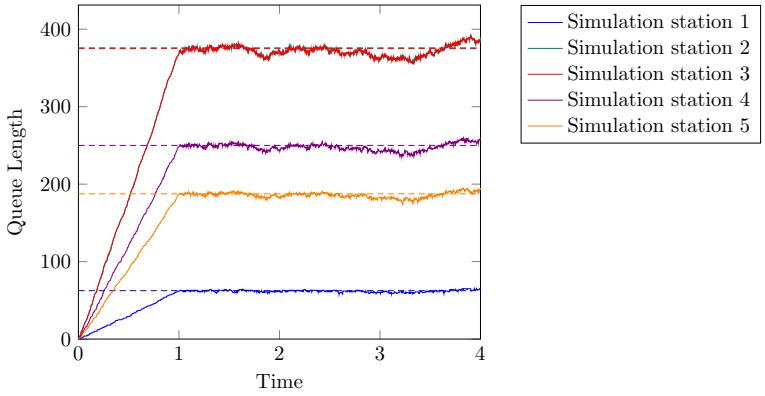

Fig. 7 Sample paths of the queue lengths for charging times equal to one when $Q^r(0) = (0, 0, 0, 0, 0)$

asymptotic expressions. This implies that the provisioning rules (2) guarantee that a desired waiting probability is achieved.

5.2 Universality result for charging time distribution

In order to be able to rigorously prove the state-space and consequential results, we assumed exponential charging times in our framework. Yet, extensive simulation experiments suggest that these results hold for any charging time distribution with finite mean and variance. In Figs. 7 and 8, we consider the system setting as described in Sect. 5.1. It appears that similar behavior occurs on the fluid scale in the case of deterministic and uniformly distributed charging times as for the exponential case.

When the queue lengths are initially zero, the system behaves close to its invariant state for $t \geq 1$. Similarly to the setting with exponential charging times, the maximum difference between the diffusion-scaled queue length behaves quite erratically; see Fig. 9. Still, the differences are very small, and grow smaller as r grows larger, sug-

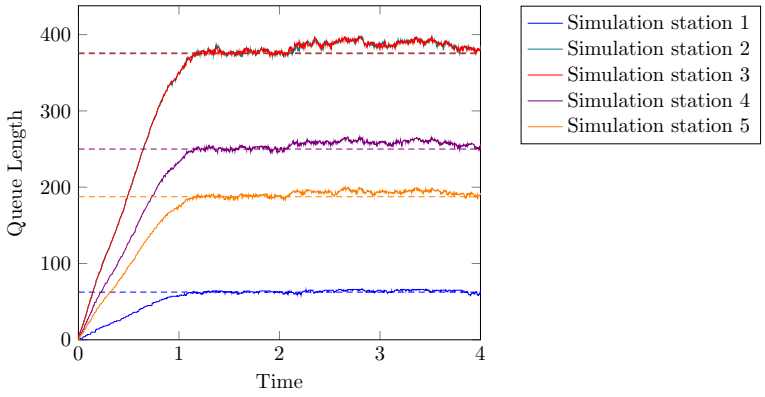

Fig. 8 Sample paths of the queue lengths for charging distribution uniform $U(0.75, 1.25)$ when $Q^r(0) = (0, 0, 0, 0, 0)$

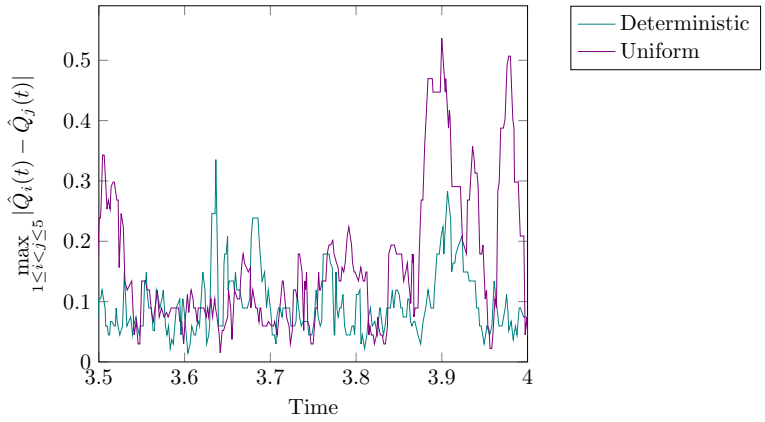

Fig. 9 Maximum distance between queue lengths for non-exponential charging times

gesting that state-space collapse also holds in this setting. That is, the system behaves similarly to the situation when there is a single station with an aggregated number of charging points and spare batteries, and a charging time distribution as at the individual stations. Consequently, performance measures such as waiting probability and expected waiting time are approximately equal to their equivalents in a single-station system.

5.3 The role of system size

In the previous sections, we commented that the differences between the diffusion-scaled queue lengths are small and fluctuate erratically among each other when one would zoom in on this domain. Obviously, the differences between the diffusion-scaled queue lengths are not arbitrarily small since r is finite. Even if the diffusion-scaled queue lengths at all stations are the same, and an arriving EV moves to station 1,

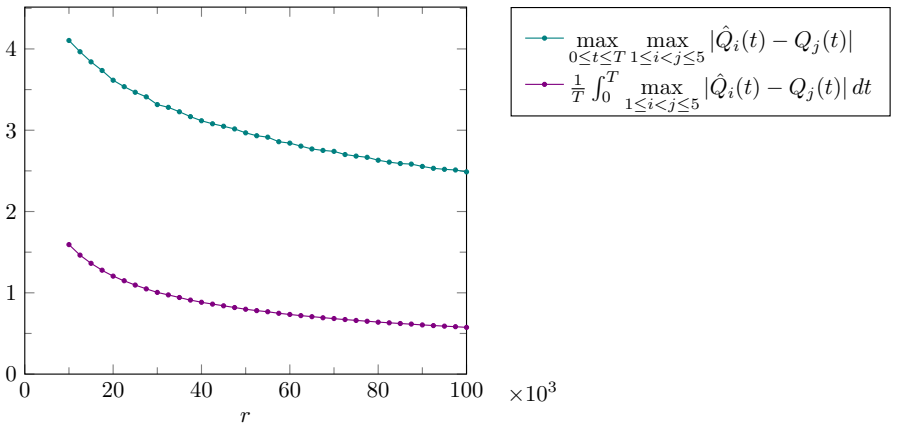

Fig. 10 Maximum queue length measures for $T = 1$ averaged over 10,000 samples

this causes a discrepancy of $1/(p_1\sqrt{\lambda r/\mu}) \approx 0.5657$ in the described setting in Sect. 5.1. Theorem 6 implies that the distance between the queue lengths become smaller as the number of EV users r grows large. To illustrate this notion, we consider the maximum difference between the queue lengths over a finite interval $T = 1$ in Fig. 10, which is monotonically decreasing in r. In addition, we observe that the average maximum distance, i.e., $1/T \int_0^T \max_{1\le i<j\le S}\{|Q_i(t) - Q_j(t)|\}\mathrm{d}t$, is also monotonically decreasing in r, and is not excessively smaller than the maximum distance of the interval.

Summarizing, as the system size increases, the accuracy of the approximations improve. However, one can imagine that a real-life battery swapping infrastructure is not of the scale as discussed in the previous sections. Therefore, we consider how our results hold up in a more realistic setting for an EV battery swapping infrastructure.

5.4 Moderate-sized system

We consider the following setting: We have the same network structure as given in Fig. 3, but with different arrival probabilities. More specifically, we assume $p_{12} = p_{23} = p_{24} = p_{25} = 0.2$ and $p_{13} = p_{34} = 0.1$, giving an effective arrival probability of $(p_1, \ldots, p_5) = (0.1, 0.25, 0.3, 0.2, 0.15)$. For the other parameters, we assume that recharging takes 4 h on average ($\mu = 0.25$) and every EV user returns for recharging services after every 50 h on average ($\lambda = 0.02$). We assume that there our infrastructure consists of a thousand electric vehicles ($r = 1000$). The effective load at the stations is

$$\left(\frac{p_j \lambda r}{\mu}\right)_{j=1,\ldots,5} = (8, 20, 24, 16, 12),$$

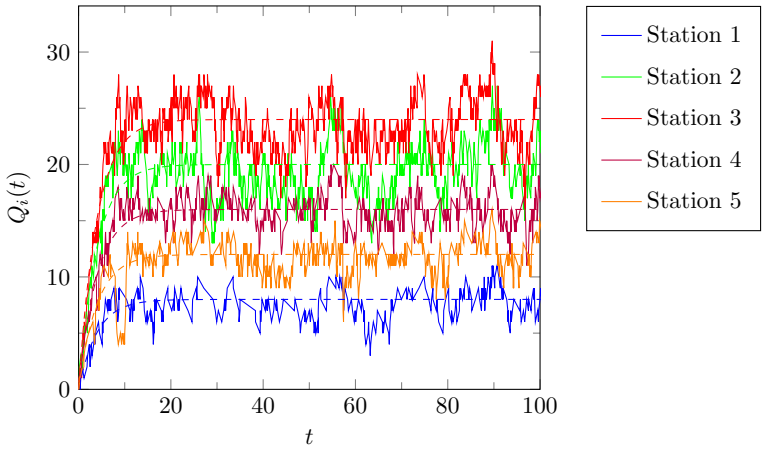

Fig. 11 Queue length behavior for moderate-sized system with $\beta = \gamma\sqrt{5}$

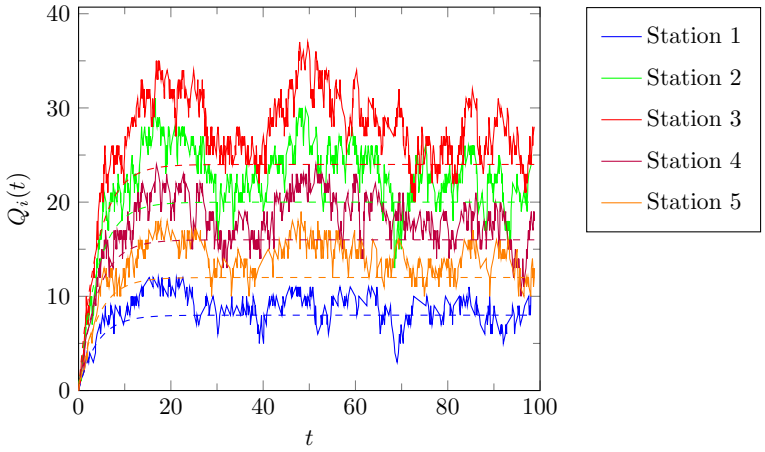

Fig. 12 Queue length behavior for moderate-sized system with $\beta = \sqrt{5}$ and $\gamma = 0$

Note that $\lambda r/\mu = 80$ and $\sqrt{\lambda r/\mu} = 4\sqrt{5} \approx 8.94$. Relatively, their sizes are much closer to one another than when r becomes larger, and this will also have its impact on the behavior.

Due to the smaller system size, we can expect that the fluctuations of the queue length around its fluid limit are relatively larger. Indeed, if one plots a sample path for this system with initial queue length $Q(0) = (0, 0, 0, 0, 0)$, we observe that the fluctuations compared to its corresponding fluid limit approximation are significant; see Figs. 11 and 12. Moreover, in Fig. 12, the queue lengths even seem to lie above the fluid limit results. We point out that this is a consequence of the high waiting probabilities for these parameter settings. Moreover, since the fluctuations are of a significant size (relatively), this leads to sample paths that appear quite far off (above) its fluid limit approximation at first glance.

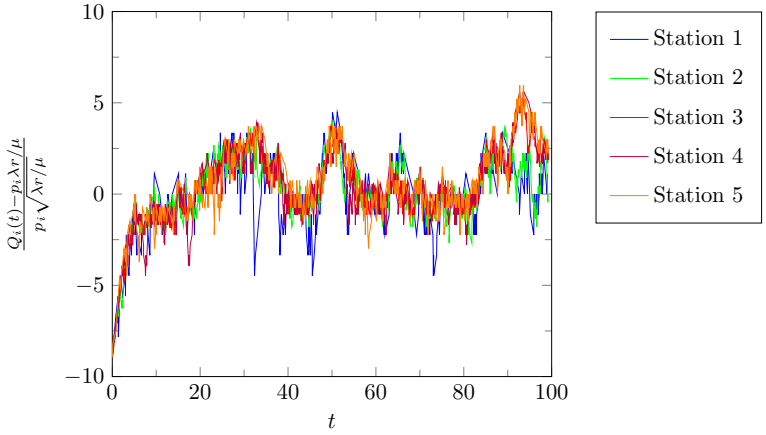

Fig. 13 Queue length behavior for moderate-sized system with $\beta = \sqrt{5}$ and $\gamma = 0$

To see whether a load-balancing effect still takes place for a moderate-sized system, we should consider the diffusion scaled queue lengths; see Fig. 13. Numerous experiments, including the setting as in Fig. 13, suggest that even for these settings, the effect of state-space collapse is very much visible. In other words, there is still a strong load-balancing effect present that leads to the occupation level at the different stations staying close to one another. In turn, this leads to performance measures, for example, waiting probability and waiting times, that are comparable at the different stations at all times.

Acknowledgements This research is supported by the Netherlands Organisation for Scientific Research through the programs Gravitation NETWORKS Grant 024.002.003, VICI Grant 639.033.413 and MEERVOUD Grant 632.003.002. The work of Seva Shneer is supported by the Mathematical Center in Akademgorodok under Agreement No. 075-15-2019-1675 with the Ministry of Science and Higher Education of the Russian Federation.

A Fluid limit proof

The proof of Theorem 5 is similar to the proof of [9], but adapted appropriately to our system.

Proof of Theorem 5 First, we show that the fluid limits exist, where all components are Lipschitz continuous. For this purpose, we show that for all $\omega \in \mathcal{A}$ the sequence

$\{\bar{\mathbb{X}}(\cdot,\omega)\}$ has a convergent subsequence, where every component in the limiting process is Lipschitz continuous.

Fix $\omega \in \mathcal{A}$. We observe that, for every $0 \le t_1 < t_2$,

$$\left|\frac{T_j^r(t_2,\omega)}{r} - \frac{T_j^r(t_1,\omega)}{r}\right| \le \frac{F_j^r(t_2-t_1)}{r} < \left(\frac{\lambda}{\mu} + |\gamma|\sqrt{\frac{\lambda}{\mu}}\right)(t_2-t_1).$$

Therefore, there exists a subsequence $\{r_l\}$ such that $\bar{T}_j^{r_l}(\cdot,\omega)$ converges u.o.c. to some $\bar{T}_j(\cdot,\omega)$ as $l \to \infty$ for every $j = 1,\ldots,S$, which is Lipschitz continuous.

Using Lemma 11 in [1], Equation (18) and the fact that $\omega \in \mathcal{A}$, it follows that $\bar{D}_j^r(\cdot,\omega)$ also converges u.o.c. to $D_j(\cdot,\omega)$ for every $j = 1,\ldots,S$. In fact, it follows that $D_j(\cdot,\omega) = \mu\bar{T}_j(\cdot,\omega)$, and is therefore also Lipschitz continuous.

Next, we consider the arrival processes. First, we observe that $L^r(t) \le r$ for every $t \ge 0$. Therefore,

$$\left|\frac{Y^r(t_2,\omega)}{r} - \frac{Y^r(t_1,\omega)}{r}\right| \le t_2 - t_1,$$

for all $0 \le t_1 < t_2$, and hence, there exists a subsequence $\{r_l\}$ such that $\bar{Y}^{r_l}(\cdot,\omega)$ converges u.o.c. to some $\bar{Y}(\cdot,\omega)$ as $l \to \infty$ for every $j = 1,\ldots,S$, which is again Lipschitz continuous.

Moreover, it follows from (16) that, for all $0 \le t_1 < t_2$,

$$\bar{A}_{ij}^r(t_2,\omega) - \bar{A}_{ij}^r(t_1,\omega) \le \frac{\Lambda_{ij}(rt_2) - \Lambda_{ij}(rt_1)}{r}.$$

As $\omega \in \mathcal{A}$ and the FSLLN applies, it follows from Theorem 12.3 in [3] that there is some subsequence $\{r_l\}$ such that $\bar{A}_{ij}^r(\cdot,\omega)$ converges u.o.c. as $l \to \infty$ to some process $\bar{A}_{ij}(\cdot,\omega)$. In particular, it holds for all $0 \le t_1 < t_2$ that

$$\bar{A}_{ij}(t_2,\omega) - \bar{A}_{ij}(t_1,\omega) \le p_{ij}\lambda(t_2-t_1),$$

and $\bar{A}_{ij}$ is hence Lipschitz continuous. Similarly, we can show the same convergence result for the processes $\bar{A}_{ij,i}^r(\cdot,\omega)$ to $\bar{A}_{ij,j}(\cdot,\omega)$.

By (17), it follows also that $\{Q_j^{r_l}(\cdot,\omega)\}$ is precompact, which in turn implies that $\{Z_j^{r_l}(\cdot,\omega)\}$ is precompact due to (21). Moreover, $\{L^{r_l}(\cdot,\omega)\}$ is precompact by (22). In conclusion, the fluid limit exists with each component being Lipschitz continuous.

Fluid equations (24)–(31) follow from the FSLLN results and applying Lemma 11 of [1]. Equation (32) requires additional arguments. Suppose $\bar{\mathbb{X}}$ to be a fluid limit with corresponding $\omega \in \mathcal{A}$ and subsequence $\{r_l\}_{l\in\mathbb{N}}$. If, for some $t > 0$, we have that $\bar{Q}_j(t)/p_j > \bar{Q}_i(t)/p_i$, then it follows by the continuity of the fluid limit that there exists a $\delta > 0$ such that

$$\frac{\bar{Q}_j(s)}{p_j} > \frac{\bar{Q}_i(s)}{p_i}$$

for all $s \in [t-\delta, t+\delta]$. By the definition of the fluid limit, it holds for large enough r_l that

$$\frac{\bar{Q}_j^{r_l}(s,\omega)}{p_j} > \frac{\bar{Q}_i^{r_l}(s,\omega)}{p_i}$$

for all $s \in [t-\delta, t+\delta]$. In this case, the routing policy states that all arrivals of type $\{i, j\}$ move to station i. Therefore, $A_{ij,j}^{r_l}$ remains constant on $[t-\delta, t+\delta]$, and hence, $\bar{A}'_{ij,j}(t) = 0$. Moreover, station i receives all arrivals and, by the FSLLN and (25),

$$\bar{A}_{ij,i}(t_2,\omega) - \bar{A}_{ij,i}(t_1,\omega) = p_{ij}\lambda\bar{L}(t,\omega)(t_2 - t_1)$$

for all $t_1 < t_2$ with $t_1, t_2 \in [t-\delta, t+\delta]$. It follows that $\bar{A}'_{ij,i}(t,\omega) = p_{ij}\lambda\bar{L}(t,\omega)$ by (15).

Finally, we show that there is a unique invariant state given by $q = (p_1\lambda/\mu, \ldots, p_S\lambda/\mu)$. Introduce the function

$$h(t) = \max_{1\le j\le S}\left\{\frac{\bar{Q}_j(t)}{p_j}\right\} - \min_{1\le j\le S}\left\{\frac{\bar{Q}_j(t)}{p_j}\right\},$$

and write

$$\bar{S}_{\max}(t) = \arg\max_{1\le j\le S}\left\{\frac{\bar{Q}_j(t)}{p_j}\right\}, \quad \bar{S}_{\min}(t) = \arg\min_{1\le j\le S}\left\{\frac{\bar{Q}_j(t)}{p_j}\right\}.$$

Trivially, $h(t) \ge 0$. Since $\bar{Q}(\cdot)$ is Lipschitz continuous, so is $h(\cdot)$, and hence, it is differentiable almost everywhere. To show that $h(0) = 0$ implies $h(t) = 0$ for all $t \ge 0$, it therefore suffices to show that if $h(t) > 0$ then $h'(t) < 0$ for every regular point t of $\bar{\mathbb{X}}$. By (26), we observe that, for every $j = 1, \ldots, S$ and regular point $t \ge 0$,

$$\bar{Q}'_j(t) = \sum_{i:\{i,j\}\in E} \bar{A}'_{ij,j}(t) - \bar{D}'_j(t).$$

Due to (27)–(30), $\bar{D}'_j(t) = \min\{\mu\bar{Q}_j(t), p_j\lambda\}$. In particular, we observe that $\bar{D}'_i(t)/p_i = \bar{D}'_j(t)/p_j$ for all $i, j \in \bar{S}_{\max}(t)$, as well as for all $i, j \in \bar{S}_{\min}(t)$. Due to Lemma 2.8.6 from [8], as t is a regular point, it follows that

$$\frac{\sum_{i:\{i,j\}\in E} \bar{A}'_{ij,j}(t)}{p_j} = \frac{\sum_{k:\{k,l\}\in E} \bar{A}'_{kl,l}(t)}{p_l}$$

for every $j, l \in \bar{S}_{\max}(t)$, as well as for every $j, l \in \bar{S}_{\min}(t)$. Due to (24), (25) and (32), we conclude that, for every $j \in S_{\max}(t)$,

$$\frac{\sum_{i:\{i,j\}\in E} \bar{A}'_{ij,j}(t)}{p_j} = \frac{1}{p_j}\left(\frac{p_j}{\sum_{i\in\bar{S}_{\max}(t)} p_i} \sum_{\substack{\{i,j\}\in E,\\ i,j\in S_{\max}(t)}} p_{ij}\lambda\bar{L}(t)\right) < \lambda\bar{L}(t).$$

On the other hand, for every $j \in S_{\min}(t)$,

$$\frac{\sum_{i:\{i,j\}\in E} \bar{A}'_{ij,j}(t)}{p_j} = \frac{1}{p_j}\left(\frac{p_j}{\sum_{i\in\bar{S}_{\min}(t)} p_i} \sum_{\substack{\{i,j\}\in E,\\ i\in\bar{S}_{\min}(t)\cup j\in\bar{S}_{\min}(t)}} p_{ij}\lambda\bar{L}(t)\right) > \lambda\bar{L}(t).$$

Observing that $\bar{D}'_j(t)/p_j > \bar{D}'_i(t)/p_i$ for every $j \in S_{\max}(t)$ and $i \in S_{\min}(t)$, we therefore conclude that if $h(t) > 0$ with t a regular point, then

$$h'(t) < \lambda\bar{L}(t) - \lambda\bar{L}(t) = 0.$$

In other words, for every invariant state of the fluid limit, it must hold that $\bar{Q}_i/p_i(t) = \bar{Q}_j/p_j(t)$ for every $i \neq j$. We observe, in view of (26), that

$$\frac{\bar{Q}'_j(t)}{p_j} = \frac{\sum_{i:\{i,j\}\in E} \bar{A}'_{ij,j}(t) - \bar{D}'_j(t)}{p_j},$$

where, for every $1 \leq j \leq S$,

$$\frac{\bar{D}'_j(t)}{p_j} = \mu \min\{\bar{Q}_j(t)/p_j, \lambda/\mu\},$$

and hence, in view of (24) and (25),

$$\sum_{i:\{i,j\}\in E} \bar{A}'_{ij,j}(t) = p_j\lambda\bar{L}(t).$$

That is, $\bar{Q}'_j(t) = 0$ if and only if

$$\mu \min\{\bar{Q}_j(t)/p_j, \lambda/\mu\} = p_j\lambda\bar{L}(t).$$

In view of (31), this occurs if and only if $\bar{Q}_j(t) = p_j\lambda/\mu$ for every $j = 1, \ldots, S$.

□

B State-space collapse proofs

To prove Theorem 6, we use a framework similar to that of [6,9]. The construction consists of several steps, which we lay out next.

1. Divide the interval $[0, T]$ into $T\sqrt{r}$ intervals of length $1/\sqrt{r}$, indexed by m. In each interval, consider the *hydrodynamically* scaled process of $\mathbb{X}$. For each of these intervals, we
   (a) show the scaled process is "almost" Lipschitz continuous;
   (b) show convergence to some *hydrodynamic* limiting process for a sufficiently large part of the state space;
   (c) derive the hydrodynamic limit equations.
2. Relate the hydrodynamic scaling to the diffusion scaling, using a SSC function to deal with complications regarding the range of the time variable. Transferring the results appropriately, we show *multiplicative* SSC with respect to the SSC function.
3. Using a compact containment condition, we show that this implies *strong* SSC.

B.1 Hydrodynamic scaling and its limiting process

In order to introduce the hydrodynamic scaling, we use a diffusion scaling for the values of the process but we slow the process down in time in order to analyze what occurs initially (what would happen instantaneously on a diffusive scale). That is, we divide the interval $[0, T]$ in $T\sqrt{r}$ intervals of length $1/\sqrt{r}$, indexed by m. We write $p = (p_1, \ldots, p_S)$, and

$$x_{r,m} = \max\left\{\left|Q^r\left(\frac{m}{\sqrt{r}}\right) - p\frac{\lambda r}{\mu}\right|^2, \left|L^r\left(\frac{m}{\sqrt{r}}\right) - r\right|^2, r\right\}. \tag{37}$$

For the processes in $\mathbb{X}$, we introduce the following hydrodynamically scaled variants: For Q^r, Z^r and L^r, let

$$Q^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(Q^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - p\frac{\lambda r}{\mu}\right), \tag{38}$$

$$Z^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(Z^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - p\frac{\lambda r}{\mu}\right), \tag{39}$$

$$L^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(L^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - r\right), \tag{40}$$

the deviations of these processes with respect to their fluid limits. For the processes A^r, A_d, Y^r, T^r and D_r, we introduce

$$A^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(A^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - A^r\left(\frac{m}{\sqrt{r}}\right)\right), \tag{41}$$

$$A_d^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(A_d^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - A_d^r\left(\frac{m}{\sqrt{r}}\right)\right), \tag{42}$$

$$Y^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(Y^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - Y^r\left(\frac{m}{\sqrt{r}}\right)\right), \tag{43}$$

$$T^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(T^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - T^r\left(\frac{m}{\sqrt{r}}\right)\right), \tag{44}$$

$$D^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}}\left(D^r\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) - D^r\left(\frac{m}{\sqrt{r}}\right)\right). \tag{45}$$

In other words, we track the increase of these processes during the interval $[m/\sqrt{r}, m/\sqrt{r} + \sqrt{x_{r,m}}t/r]$. By the definition of $x_{r,m}$, we note that

$$|\mathbb{X}^{r,m}(0)| \leq 1,$$

which will be a required compactness property when we prove convergence to a hydrodynamic limit. Moreover, due to our fluid limit results, we can show that $\sqrt{x_{r,m}}/r$ is very small for all $\omega \in \mathcal{A}$.

Lemma 2 *Suppose $\hat{Q}^r(0) \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$, and let $M > 0$ be fixed. For every $\epsilon > 0$ and $\omega \in \mathcal{A}$,*

$$\max_{m<\sqrt{r}T}\left\{\frac{\sqrt{x_{r,m}}}{r}\|Q^{r,m}(t)\|_M, \frac{\sqrt{x_{r,m}}}{r}\|L^{r,m}(t)\|_M\right\} \leq \epsilon$$

for r large enough.

Proof Due to our fluid limit result in Theorem 5 and the definition of $x_{r,m}$ in (37), we observe that

$$\frac{\sqrt{x_{r,m}}}{r} \leq \max\{\|Q^r(t) - p\lambda r/\mu\|_T/r, \|L^r(t) - r\|_T/r, 1/\sqrt{r}\} \leq \epsilon$$

for r large enough. Moreover, for r large enough,

$$\max\left\{\left\|Q^r(t) - p\lambda r/\mu\right\|_{T+M\epsilon}/r, \|L^r(t) - r\|_{T+M\epsilon}/r\right\} \leq \epsilon.$$

We conclude that, for every $m \leq \sqrt{r}T$,

$$\frac{\sqrt{x_{r,m}}}{r}\|Q^{r,m}(t)\|_M = \frac{\|Q^r(m/\sqrt{r} + \sqrt{x_{r,m}}/rt) - p\lambda r/\mu\|_M}{r} \leq \epsilon,$$

and

$$\frac{\sqrt{x_{r,m}}}{r}\|L^{r,m}(t) - r\|_M \leq \epsilon.$$

□

For the hydrodynamically scaled process, the system identities translate to

$$A_{ij}^{r,m}(t) = A_{ij,i}^{r,m}(t) + A_{ij,j}^{r,m}(t), \quad \forall \{i,j\} \in E, \tag{46}$$

$$A_{ij}^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}} \left(\Lambda_{ij} \left(Y^r(m/\sqrt{r}) + \sqrt{x_{r,m}} Y^{r,m}(t) \right) - \Lambda_{ij} \left(Y^r(m/\sqrt{r}) \right) \right), \quad \forall \{i,j\} \in E, \tag{47}$$

$$Q_j^{r,m}(t) = Q_j^{r,m}(0) + \sum_{i:\{i,j\}\in E} A_{ij,j}^{r,m}(t) - D_j^{r,m}(t), \quad \forall j = 1, \ldots, S, \tag{48}$$

$$D_j^{r,m}(t) = \frac{1}{\sqrt{x_{r,m}}} \left(S_j \left(T_j^r(m/\sqrt{r}) + \sqrt{x_{r,m}} T_j^{r,m}(t) \right) - S_j \left(T_j^r(m/\sqrt{r}) \right) \right), \quad \forall j = 1, \ldots, S, \tag{49}$$

$$Y^{r,m}(t) = t + \frac{\sqrt{x_{r,m}}}{r} \int_0^t L^{r,m}(s)\, \mathrm{d}s, \quad \forall j = 1, \ldots, S, \tag{50}$$

$$T_j^{r,m}(t) = p_j \frac{\lambda}{\mu} t + \frac{\sqrt{x_{r,m}}}{r} \int_0^t Z_j^{r,m}(s)\, \mathrm{d}s, \quad \forall j = 1, \ldots, S, \tag{51}$$

$$Z_j^{r,m}(t) = \min \left\{ Q_j^{r,m}(t), \frac{1}{\sqrt{x_{r,m}}} \gamma p \sqrt{\frac{\lambda r}{\mu}} \right\}, \quad \forall j = 1, \ldots, S, \tag{52}$$

$$L^{r,m}(t) = -\sum_{j=1}^{S} \left(Q_j^{r,m}(t) - \frac{1}{\sqrt{x_{r,m}}} \beta p_j \sqrt{\frac{\lambda r}{\mu}} \right)^+, \tag{53}$$

$$A_{ij,i}^{r,m}(t) \text{ can only increase when } \frac{Q_i^{r,m}(t)}{p_i} \le \frac{Q_j^{r,m}(t)}{p_j} \quad \forall \{i,j\} \in E. \tag{54}$$

In order to show that $\mathbb{X}^{r,m}$ is almost (with the exception of certain events) Lipschitz continuous, we would like to exclude these certain events, i.e., show that such events are unlikely to occur.

Lemma 3 *Fix $\epsilon > 0$, $M > 0$ and $T > 0$. For r large enough, there exists a constant $N > 0$ (only depending on λ, μ, and $\{p_{ij}; \{i,j\} \in E\}$) such that*

$$\mathbb{P}\left(\max_{m<\sqrt{r}T} \sup_{0\le t_1\le t_2\le M} \left\{ \left| A^{r,m}(t_2) - A^{r,m}(t_1) \right| - N(t_2 - t_1) \right\} \ge \epsilon \right) \le \epsilon, \tag{55}$$

$$\mathbb{P}\left(\max_{m<\sqrt{r}T} \sup_{0\le t_1\le t_2\le M} \left\{ \left| D^{r,m}(t_2) - D^{r,m}(t_1) \right| - N(t_2 - t_1) \right\} \ge \epsilon \right) \le \epsilon. \tag{56}$$

Moreover,

$$\mathbb{P}\left(\max_{m<\sqrt{r}T} \left\| Y^{r,m}(t) - \frac{1}{p_{ij}\lambda} A_{ij}^{r,m}(t) \right\|_M \ge \epsilon \right) \le \epsilon, \quad \{i,j\} \in E, \tag{57}$$

$$\mathbb{P}\left(\max_{m<\sqrt{r}T} \left\| T_j^{r,m}(t) - \frac{1}{\mu} D_j^{r,m}(t) \right\|_M \ge \epsilon \right) \le \epsilon, \quad j = 1, \ldots, S. \tag{58}$$

Proof First, we note that, due to the memoryless property which both the arrival and service completion processes satisfy, the choice of m is irrelevant and thus can be made arbitrarily. To prove (55), we start by showing that, for every $\{i, j\} \in E$,

$$\mathbb{P}\left(\sup_{0\le t_1\le t_2\le M}\left\{A_{ij}^{r,m}(t_2)-A_{ij}^{r,m}(t_1)-\frac{1}{p_{ij}\lambda}(t_2-t_1)\right\}\ge\epsilon\right)\le\epsilon.$$

From (50) and (53), we observe that $Y_{r,m}(t) \le t$ and is non-decreasing. Due to the properties of Poisson processes,

$$\begin{aligned}A_{ij}^{r,m}(t_2)-A_{ij}^{r,m}(t_1)&\stackrel{d}{=}\frac{\Lambda_{ij}\left(\sqrt{x_{r,m}}Y^{r,m}(t_2)\right)-\Lambda_{ij}\left(\sqrt{x_{r,m}}Y^{r,m}(t_1)\right)}{\sqrt{x_{r,m}}}\\&\stackrel{d}{\le}\frac{\Lambda_{ij}\left(\sqrt{x_{r,m}}t_2\right)-\Lambda_{ij}\left(\sqrt{x_{r,m}}t_1\right)}{\sqrt{x_{r,m}}}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}\left(\sup_{0\le t_1\le t_2\le M}\left\{A_{ij}^{r,m}(t_2)-A_{ij}^{r,m}(t_1)-\frac{t_2-t_1}{p_{ij}\lambda}\right\}\ge\epsilon\right)\\&\quad\le\mathbb{P}\left(\sup_{0\le t_1\le t_2\le M}\left\{\frac{\Lambda_{ij}\left(\sqrt{x_{r,m}}t_2\right)-\Lambda_{ij}\left(\sqrt{x_{r,m}}t_1\right)}{\sqrt{x_{r,m}}}-\frac{t_2-t_1}{p_{ij}\lambda}\right\}\ge\epsilon\right)\\&\quad\le\mathbb{P}\left(\left\|\frac{\Lambda_{ij}\left(\sqrt{x_{r,m}}t\right)}{\sqrt{x_{r,m}}}-\frac{t}{p_{ij}\lambda}\right\|_M\ge\epsilon/2\right)\le\frac{\epsilon}{2M^2\sqrt{x_{r,m}}}\le\frac{\epsilon}{2M^2\sqrt{r}},\end{aligned}$$

where the second-to-last inequality follows from Proposition 4.3 in [6]. Choosing $N = 1/(\lambda \min_{\{i,j\}\in E} p_{ij})$ and applying the union bound twice, we obtain

$$\mathbb{P}\left(\max_{m<\sqrt{r}T}\sup_{0\le t_1\le t_2\le M}\left\{\left|A^{r,m}(t_2)-A^{r,m}(t_1)\right|-N(t_2-t_1)\right\}\ge\epsilon\right)\le\frac{\epsilon T|E|}{2M^2},$$

which yields (55).

The proof for (56) is completely analogous, but with minor adaptions as one uses $T_j^{r,m}(t) \le p_j\lambda/\mu t$ instead of $Y^{r,m}(t) \le t$. We conclude that (55) and (56) show that the hydrodynamically scaled arrival process and service completion process are almost Lipschitz continuous.

In order to prove (57) and (58), we introduce the following processes: Let $\{u_{ij}(l), l \ge 1\}$ be independent exponentially distributed random variables with rate $p_{ij}\lambda$, representing the time that a car has before it needs recharging at either station i or j. Let $\{v_j(l), l \ge 1\}$ be independent exponentially distributed random variables with rate μ, representing the service requirement (recharging time) of a battery at station j. Define

$$U_{ij}(n) = \sum_{l=1}^{n} u_{ij}(l), \quad \{i, j\} \in E,$$
$$V_j(n) = \sum_{l=1}^{n} v_j(l), \quad j = 1, \ldots, S,$$

the aggregated interarrival time of n cars that will choose between stations i and j, and the total service requirement of n batteries at station j, respectively. We observe the identities

$$\Lambda_{ij}(t) = \max\{n : U_{ij}(n) \le t\}, \quad S_j(t) = \max\{n : V_j(n) \le t\}, \quad t \ge 0.$$

Moreover, due to (16) and (18), we observe

$$U_{ij}(A^r_{ij}(t)) \le Y^r(t) \le U_{ij}(A^r_{ij}(t) + 1), \qquad \{i, j\} \in E, \tag{59}$$
$$V_j(D^r_j(t)) \le T^r_j(t) \le V_j(D_j(t) + 1), \quad j = 1, \ldots, S. \tag{60}$$

As in [9], we define for notational convention, for $b = (b_1, b_2) \in \mathbb{N}$,

$$U^{r,m}_{ij}\left(A^{r,m}_{ij}(t), b\right) = \frac{1}{\sqrt{x_{r,m}}}\left(U^r_{ij}\left(A^r_{ij}\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) + b_1\right) - U^r_{ij}\left(A^r_{ij}\left(\frac{m}{\sqrt{r}}\right) + b_2\right)\right), \tag{61}$$
$$V^{r,m}_{j}\left(D^{r,m}_{j}(t), b\right) = \frac{1}{\sqrt{x_{r,m}}}\left(V^r_j\left(D^r_j\left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r}\right) + b_1\right) - V^r_j\left(D^r_j\left(\frac{m}{\sqrt{r}}\right) + b_2\right)\right). \tag{62}$$

In view of (59) and (60), this yields the inequalities

$$U^{r,m}_{ij}(A^r_{ij}(t), (0, 1)) \le Y^{r,m}(t) \le U^{r,m}_{ij}(A^r_{ij}(t), (1, 0)), \quad \{i, j\} \in E, \tag{63}$$
$$V^{r,m}_{j}(D_j(t), (0, 1)) \le T^{r,m}_{j}(t) \le V^{r,m}_{j}(D_j(t), (1, 0)), \quad j = 1, \ldots, S. \tag{64}$$

Using these processes, we first prove the following bounds:

$$\mathbb{P}\left(\max_{m<\sqrt{r}T} \left\| U_{ij}(A^{r,m}_{ij}(t), b) - \frac{1}{p_{ij}\lambda} A^{r,m}_{ij}(t) \right\|_M \ge \epsilon\right) \le \epsilon, \quad \forall \{i, j\} \in E, \tag{65}$$
$$\mathbb{P}\left(\max_{m<\sqrt{r}T} \left\| V_j(D^{r,m}(t), b) - \frac{1}{\mu} D^{r,m}_j(t) \right\|_M \ge \epsilon\right) \le \epsilon, \quad j = 1, \ldots, S, \tag{66}$$

for $b = (1, 0)$ and $b = (0, 1)$. The proof is similar to that of (78) in [21]. We observe that the proof of (55) implies that in particular

$$\mathbb{P}\left(A^r_{ij}\left(\frac{\sqrt{x_{r,m}}M}{r}\right) \ge \frac{2M}{p_{ij}\lambda}\sqrt{x_{r,m}}\right) \le \frac{\epsilon}{M^2\sqrt{r}},$$

and hence also

$$\mathbb{P}\left(A_{ij}^r\left(\frac{\sqrt{x_{r,m}}M}{r}\right)+1\geq\frac{3M}{p_{ij}\lambda}\sqrt{x_{r,m}}\right)\leq\frac{\epsilon}{M^2\sqrt{r}}$$

for r large enough. Proposition 4.2 of [6] states

$$\mathbb{P}\left(\left\|U_{ij}(l)-\frac{l}{p_{ij}\lambda}\right\|_n\geq\epsilon n\right)\leq\frac{\epsilon}{n}.$$

Therefore, it follows that

$$\mathbb{P}\left(\left\|U_{ij}(A_{ij}^r(t))-\frac{A_{ij}^r(t)}{p_{ij}\lambda}\right\|_{\sqrt{x_{r,m}}M/r}\geq\epsilon\frac{2M\sqrt{x_{r,m}}}{p_{ij}\lambda}\right)\leq\frac{\epsilon}{\sqrt{r}}\left(\frac{1}{M^2}+\frac{p_{ij}\lambda}{2M}\right),$$

and

$$\mathbb{P}\left(\left\|U_{ij}(A_{ij}^r(t)+1)-\frac{A_{ij}^r(t)}{p_{ij}\lambda}\right\|_{\sqrt{x_{r,m}}M/r}\geq\epsilon\frac{3M\sqrt{x_{r,m}}}{p_{ij}\lambda}\right)\leq\frac{\epsilon}{\sqrt{r}}\left(\frac{1}{M^2}+\frac{p_{ij}\lambda}{3M}\right).$$

Increasing ϵ appropriately, we obtain

$$\mathbb{P}\left(\left\|U_{ij}^{r,m}(A_{ij}^{r,m}(t),b)-\frac{A_{ij}^{r,m}(t)}{p_{ij}\lambda}\right\|_M\geq\epsilon\right)\leq\frac{\epsilon}{T\sqrt{r}}$$

for both $b=(0,0)$ and $b=(1,0)$. Using the union bound yields

$$\mathbb{P}\left(\max_{m<\sqrt{r}T}\left\|U_{ij}^{r,m}(A_{ij}^{r,m}(t),b)-\frac{A_{ij}^{r,m}(t)}{p_{ij}\lambda}\right\|_M\geq\epsilon\right)\leq\epsilon$$

for both $b=(0,0)$ and $b=(1,0)$. To conclude the proof for $b=(0,1)$ as well, we observe

$$\begin{aligned}
&\mathbb{P}\left(\max_{m<\sqrt{r}T}\left\|U_{ij}^{r,m}(A_{ij}^{r,m}(t),b)-U_{ij}^{r,m}(A_{ij}^{r,m}(t),(0,0))\right\|_M\geq\epsilon\right)\\
&\leq\mathbb{P}\left(\max_{m<\sqrt{r}T}\left|U_{ij}\left(A_{ij}^r\left(\frac{m}{\sqrt{r}}\right)+1\right)-U_{ij}\left(A_{ij}^r\left(\frac{m}{\sqrt{r}}\right)\right)\right|\geq\epsilon\sqrt{x_{r,m}}\right)\\
&\leq\mathbb{P}\left(u_{ij}^{r,T,\max}\geq\sqrt{x_{r,m}}\epsilon\right)\leq\epsilon,
\end{aligned}$$

where the final inequality follows from Lemma 5.1 in [6] with

$$u_{ij}^{r,T,\max}=\max\{u_{ij}(l):U_i(l-1)\leq rT\}.$$

The proof of (66) is analogous to (65), replacing the arrival processes by the service processes. Equations (57) and (58) are then a direct consequence of (63) and (64). □

Using the previous result, we can show that $\mathbb{X}$ is almost Lipschitz continuous.

Proposition 2 *Fix $\epsilon > 0$, $M > 0$ and $T > 0$. For r large enough,*

$$\mathbb{P}\left(\max_{m<\sqrt{r}T}\sup_{0\le t_1\le t_2\le M}\left\{\left|\mathbb{X}^{r,m}(t_2)-\mathbb{X}^{r,m}(t_1)\right|-N(t_2-t_1)\right\}\ge\epsilon\right)\le\epsilon,$$

where $N<\infty$ is constant (depending only on λ, μ and $\{p_{ij};\{i,j\}\in E\}$).

Proof This follows in a straightforward way from Lemma 3 and the hydrodynamically scaled system equations. That is, let $\mathcal{V}^r$ denote the intersection of the complements of the events given in Eqs. (55)–(58), so $\mathbb{P}(\mathcal{V}^r)\le 1-N_0\epsilon$ with N_0 the number of equations in Lemma 3. We note that in order to prove the proposition, it suffices to show that for every $\omega\in\mathcal{V}^r$, and for every $t_1,t_2\in[0,T]$ and $m<\sqrt{r}T$,

$$\left|\mathbb{X}^{r,m}(t_2)-\mathbb{X}^{r,m}(t_1)\right|\le N_1(t_2-t_1)+N_2\epsilon, \tag{67}$$

where N_1 and N_2 are only dependent on the system parameters (i.e., λ, μ, p). Let $t_1,t_2\in[0,T]$ with $t_1\le t_2$. By the definition of $\mathcal{V}^r$,

$$\left|A^{r,m}(t_2)-A^{r,m}(t_1)\right|\le N(t_2-t_1)+\epsilon,$$

and

$$\left|D^{r,m}(t_2)-D^{r,m}(t_1)\right|\le N(t_2-t_1)+\epsilon,$$

for N as in Lemma 3. Due to (46),

$$\left|A_d^{r,m}(t_2)-A_d^{r,m}(t_1)\right|\le\left|A^{r,m}(t_2)-A^{r,m}(t_1)\right|\le N(t_2-t_1)+\epsilon.$$

In view of (48) and (46), we observe

$$\begin{aligned}\left|Q^{r,m}(t_2)-Q^{r,m}(t_1)\right|&\le|E|\left|A^{r,m}(t_2)-A^{r,m}(t_1)\right|+S\left|D^{r,m}(t_2)-D^{r,m}(t_1)\right|\\&\le(|E|+S)N(t_2-t_1)+2\epsilon.\end{aligned}$$

Due to (57),

$$\begin{aligned}\left|Y^{r,m}(t_2)-Y^{r,m}(t_1)\right|&\le\sum_{\{i,j\}\in E}\frac{|A^{r,m}(t_2)-A^{r,m}(t_1)|}{p_{ij}\lambda}+2\epsilon\\&\le\sum_{\{i,j\}\in E}\frac{N}{p_{ij}\lambda}(t_2-t_1)+\left(\sum_{\{i,j\}\in E}\frac{1}{p_{ij}\lambda}+2\right)\epsilon.\end{aligned}$$

and similarly, due to (58),

$$\left|T^{r,m}(t_2) - T^{r,m}(t_1)\right| \leq \frac{1}{\mu}\left|D^{r,m}(t_2) - D^{r,m}(t_1)\right| + 2\epsilon \leq \frac{N}{\mu}(t_2 - t_1) + \left(\frac{1}{\mu} + 2\right)\epsilon.$$

In view of (52),

$$\left|Z^{r,m}(t_2) - Z^{r,m}(t_1)\right| \leq \left|Q^{r,m}(t_2) - Q^{r,m}(t_1)\right| \leq (|E| + S)N(t_2 - t_1) + 2\epsilon.$$

Finally, due to (53),

$$\left|L^{r,m}(t_2) - L^{r,m}(t_1)\right| \leq S\left|Q^{r,m}(t_2) - Q^{r,m}(t_1)\right| \leq S(|E| + S)N(t_2 - t_1) + 2S\epsilon.$$

We conclude that (67) is satisfied, as each process in $\mathbb{X}^{r,m}$ satisfies this property.

□

As is done in [6,9], one can take ϵ appropriately small for every system. That is, for fixed $M > 0$, $N > 0$ and $T > 0$, let

$$\mathcal{K}_0^r = \left\{ \max_{m < \sqrt{r}T} \sup_{0 \leq t_1 \leq t_2 \leq M} \left|\mathbb{X}^{r,m}(t_2) - \mathbb{X}^{r,m}(t_1)\right| \geq N(t_2 - t_1) + \epsilon(r) \right\},$$

where $\epsilon(r) \to 0$ as $r \to \infty$ is a sequence of positive real numbers. Moreover, in view of Lemma 2, let, for that same sequence $\{\epsilon(r)\}_{r \in \mathbb{R}}$,

$$\mathcal{H}^r = \left\{ \max_{m < \sqrt{r}T} \left\{ \frac{\sqrt{x_{r,m}}}{r} \|Q^{r,m}(t)\|_M, \frac{\sqrt{x_{r,m}}}{r} \|L^{r,m}(t)\|_M \right\} \leq \epsilon(r) \right\}.$$

Let $\mathcal{K}^r$ denote the intersection of $\mathcal{K}_0^r$, $\mathcal{H}^r$, and the complements of the events in Lemma 3. We note that Lemmas 2, 3 and Proposition 2 continue to hold for the sequence $\epsilon(r)$ if $\epsilon(r) \to 0$ sufficiently slowly. We conclude that $\mathbb{P}(\mathcal{K}^r) \to 1$ as $r \to \infty$.

Corollary 2 *Fix $M > 0$ and choose $N > 0$ and $\epsilon(r)$ as above. Then,*

$$\lim_{r \to \infty} \mathbb{P}(\mathcal{K}^r) = 1.$$

Following the framework of [6], we can use these results to state that the hydrodynamically scaled system convergences to a hydrodynamic limit. Fix $M > 0$ and let $\tilde{E}$ be the set of right-continuous functions $x : [0, M] \to \mathbb{R}^d$ with left limits. Let

$$E' = \{x \in \tilde{E} : |x(0)| \leq 1, |x(t_2) - x(t_1)| \leq N|t_2 - t_1| \quad \forall t_1, t_2 \in [0, M]\}.$$

Moreover, we set

$$E^r = \{\mathbb{X}^{r,m}, m < \sqrt{r}T, \omega \in \mathcal{K}^r\},$$

and

$$\mathcal{E} = \{E^r, r \in \mathbb{N}\}.$$

We remark that these definitions are not related to E, the set of all possible pairs of stations where cars can move to.

Definition 2 A hydrodynamic limit of $\mathcal{E}$ is a point $x \in \tilde{E}$ such that for all $\epsilon > 0$ there exists a $r_0 \in \mathbb{N}$ so that for every $r \geq r_0$ there is some $y \in E^r$ such that $|x(\cdot) - y(\cdot)|_M < \epsilon$.

Since $|\mathbb{X}^{r,m}(0)| \leq 1$, the following result is a consequence of Proposition 4.1 in [6].

Corollary 3 *Let $\tilde{E}$, E^r, $\mathcal{E}$ be as above. Fix $\epsilon > 0$, $M > 0$, $T \geq 0$, and choose r large enough. Then, for $\omega \in \mathcal{K}^r$ and any $m < \sqrt{r}T$,*

$$\|\mathbb{X}^{r,m}(\cdot) - \mathbb{X}(\cdot)\|_M \leq \epsilon$$

for some hydrodynamic limit $\mathbb{X}(\cdot) \in E'$ of $\mathcal{E}$.

Finally, to conclude this section, we derive the equations that are satisfied by any hydrodynamic limit.

Proposition 3 *Let $M > 0$ be fixed, and let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathcal{E}$ over $[0, M]$. Then $\tilde{\mathbb{X}}$ satisfies the following equations:*

$$\tilde{A}_{ij}(t) = \tilde{A}_{ij,i}(t) + \tilde{A}_{ij,j}(t), \quad \forall \{i, j\} \in E, \tag{68}$$

$$\tilde{A}_{ij}(t) = p_{ij}\lambda \tilde{Y}(t) = p_{ij}\lambda t \quad \forall \{i, j\} \in E, \tag{69}$$

$$\tilde{Q}_j(t) = \tilde{Q}_j(0) + \sum_{i:\{i,j\}\in E} \tilde{A}_{ij,j}(t) - \tilde{D}_j(t), \quad \forall j = 1, \ldots, S, \tag{70}$$

$$\tilde{D}_j(t) = \mu \tilde{T}_j(t) = p_j \lambda t, \quad \forall j = 1, \ldots, S, \tag{71}$$

$$\tilde{Y}(t) = t, \quad \forall j = 1, \ldots, S, \tag{72}$$

$$\tilde{T}_j(t) = p_j \lambda / \mu t, \quad \forall j = 1, \ldots, S, \tag{73}$$

$$\tilde{A}'_{ij,i}(t) = \begin{cases} p_{ij}\lambda & \text{if } \frac{\tilde{Q}_i(t)}{p_i} < \frac{\tilde{Q}_j(t)}{p_j} \\ 0 & \text{if } \frac{\tilde{Q}_j(t)}{p_j} > \frac{\tilde{Q}_i(t)}{p_i} \end{cases} \quad \forall \{i, j\} \in E. \tag{74}$$

Remark 4 We cannot provide such general equations for $\tilde{Z}(\cdot)$ or $\tilde{L}(\cdot)$, since these limits depend on $x_{r,m}$. That is, the processes $\tilde{Z}^{r,m}(\cdot)$ and $\tilde{L}^{r,m}(\cdot)$ converge to a limit, but the limiting process may differ for different m. In the proof, we specify the limiting equations of these processes as well.

Proof of Proposition 3 Let $\tilde{\mathbb{X}}$ be a hydrodynamic limit of $\mathcal{E}$. For a given $\delta > 0$, choose (r, m) such that $\epsilon(r) \leq \delta$, and

$$\|\tilde{\mathbb{X}}(t) - \mathbb{X}^{r,m}(t, \omega)\|_M \leq \delta.$$

Due to (50), (51) and $\omega \in \mathcal{H}^r$, we derive

$$\|\tilde{Y}(t) - t\|_M \le (1+M)\delta, \ \|\tilde{T}_j(t) - p_j\lambda/\mu t\|_M \le (1+M)\delta, \quad j = 1, \ldots, S.$$

From (57) and (58), we obtain

$$\|\tilde{A}_{ij}(t) - p_{ij}\lambda t\|_M \le (2+M)\delta, \ \ \{i, j\} \in E,$$

and

$$\|\tilde{D}_j(t) - p_j\lambda t\|_M \le (2+M)\delta, \quad j = 1, \ldots, S.$$

Equation (68) is a clear consequence of (46). Combining the above equations, we observe

$$\left\| \tilde{Q}_j(t) - \tilde{Q}_j(0) - \sum_{i:\{i,j\}\in E} \tilde{A}_{ij,j}(t) + \tilde{D}_j(t) \right\|_M \le 2(|E| + S)(2+M)\delta.$$

These bounds imply that any hydrodynamic limit satisfies Eqs. (68)–(73). Finally, we still have to show (74). If, for some $t \in [0, M]$,

$$\frac{\tilde{Q}_j(t)}{p_j} > \frac{\tilde{Q}_i(t)}{p_i},$$

then by continuity of $\tilde{\mathbb{X}}$, there exists a $\eta > 0$ such that this holds for all $s \in [t-\eta, t+\eta)$, and also

$$\frac{\tilde{Q}_j^{r,m}(t)}{p_j} > \frac{\tilde{Q}_i^{r,m}(t)}{p_i}.$$

Due to (54), this implies that $A_{ij,i}^{r,m}(s)$ is constant on $s \in [t-\eta, t+\eta]$. Therefore, its limit is also constant on $[t-\eta, t+\eta]$, and hence, the derivative is zero. On the other hand, if

$$\frac{\tilde{Q}_j(t)}{p_j} < \frac{\tilde{Q}_i(t)}{p_i},$$

then by continuity of $\tilde{\mathbb{X}}$, there exists a $\eta > 0$ such that this holds for all $s \in [t-\eta, t+\eta]$, and

$$\frac{\tilde{Q}_j^{r,m}(t)}{p_j} < \frac{\tilde{Q}_i^{r,m}(t)}{p_i}.$$

Since $\tilde{A}'_{ij,j} = 0$, and due to (46) with limiting process (69),

$$\tilde{A}'_{ij,i} = \lim_{\eta \downarrow 0} \frac{\tilde{A}'_{ij}(t+\eta) - \tilde{A}'_{ij}(t)}{\eta} = p_{ij}\lambda.$$

□

B.2 The SSC function

In this section, we introduce the state-space collapse (SSC) function under which we show multiplicative state-space collapse. The SSC function we use in our paper is $g : \mathbb{R}^S \to \mathbb{R}$, defined as

$$g(q) = \max_{1 \le j \le S} \frac{q_j}{p_j} - \min_{1 \le j \le S} \frac{q_j}{p_j}, \tag{75}$$

where $q = (q_1, \ldots, q_S)$. We note that $g(\cdot)$ is a non-negative continuous function and satisfies

$$g(\alpha q) = \alpha g(q)$$

for every $\alpha > 0$.

Lemma 4 *Suppose $g : \mathbb{R}^S \to \mathbb{R}$ is defined as in (75). Then,*

$$g(\tilde{Q}(t)) \le H(t), \quad \forall t \ge 0,$$

for every hydrodynamic model solution $\tilde{\mathbb{X}}$ satisfying $|\tilde{\mathbb{X}}(0)| \le 1$, where

$$H(t) = \left(\frac{2}{\min_{1 \le j \le S} p_j} - ht \right)^+ \tag{76}$$

with $h > 0$ some constant that depends only on λ and $\{p_{ij}, \{i, j\} \in E\}$. Moreover, if $g(\tilde{Q}(0)) = 0$ and $|\tilde{\mathbb{X}}(0)| \le 1$, then $g(\tilde{Q}(t)) = 0$ for all $t \ge 0$.

Proof The proof relies heavily on the ideas used in the proof for the fluid limit. Let

$$h = \min \left\{ \lambda \frac{\sum_{\{i,j\} \in E,\ i \in \mathcal{I} \cup j \in \mathcal{I}} p_{ij}}{\sum_{i \in \mathcal{I}} p_i} - \lambda \frac{\sum_{\{i,j\} \in E,\ i \in \mathcal{J} \cap j \in \mathcal{J}} p_{ij}}{\sum_{j \in \mathcal{J}} p_j} : \emptyset \ne \mathcal{I}, \mathcal{J} \subset \{1, \ldots, S\}, \mathcal{I} \cap \mathcal{J} = \emptyset \right\}.$$

Since

$$\frac{1}{\sum_{i \in \mathcal{I}} p_i} \sum_{\substack{\{i,j\} \in E, \\ i \in \mathcal{I} \cup j \in \mathcal{I}}} p_{ij} > 1, \quad \frac{1}{\sum_{j \in \mathcal{J}} p_j} \sum_{\substack{\{i,j\} \in E, \\ i \in \mathcal{J} \cap j \in \mathcal{J}}} p_{ij} < 1,$$

for any non-empty $\mathcal{I}, \mathcal{J} \subset \{1, \ldots, S\}$ with $\mathcal{I} \cap \mathcal{J} = \emptyset$, we observe that $h < 0$. For a hydrodynamic limiting process $\tilde{\mathbb{X}}$, let $H_{\tilde{\mathbb{X}}}(\cdot)$ be given by

$$H_{\tilde{\mathbb{X}}}(t) = \left(g(\tilde{Q}(0)) - ht\right)^+.$$

We note that this function is non-negative, and satisfies $H_{\tilde{\mathbb{X}}}(t) = 0$ for all $t \geq g(\tilde{Q}(0))/h$.

To show that $g(\tilde{Q}(t))$ is bounded by this function, we note that it suffices to show that whenever $g(\tilde{Q}(t)) > 0$ with $t \geq 0$ being a regular point of $\tilde{\mathbb{X}}$,

$$g'(\tilde{Q}(t)) \leq -h.$$

For this purpose, let

$$S_{\max}(t) = \left\{ i \in \{1, \ldots, S\} : \tilde{Q}_i(t)/p_i = \max_{1 \leq j \leq S} \tilde{Q}_j(t)/p_j \right\},$$

and

$$S_{\min}(t) = \left\{ i \in \{1, \ldots, S\} : \tilde{Q}_i(t)/p_i = \min_{1 \leq j \leq S} \tilde{Q}_j(t)/p_j \right\}.$$

Due to Lemma 2.8.6 in [8], it holds for all $i, j \in S_{\max}(t)$ that $\tilde{Q}_i'(t)/p_i = \tilde{Q}_j'(t)/p_j$, and similarly, for all $i, j \in S_{\min}(t)$ it holds that $\tilde{Q}_i'(t)/p_i = \tilde{Q}_j'(t)/p_j$. Therefore, due to hydrodynamic limit equations (68)–(74) and the observation that there is at least one station $j \notin S_{\max}(t)$ such that $\{i, j\} \in E$, it follows that

$$\frac{\tilde{Q}_i'(t)}{p_i} < \frac{\lambda}{\sum_{i \in S_{\max}(t)} p_i} \sum_{\substack{\{i,j\} \in E, \\ i,j \in S_{\max}(t)}} p_{ij} - \lambda$$

for all $i \in S_{\max}(t)$. Similarly, for all $i \in S_{\min}$,

$$\frac{\tilde{Q}_i'(t)}{p_i} > \frac{\lambda}{\sum_{i \in S_{\min}(t)} p_i} \sum_{\substack{\{i,j\} \in E, \\ i \in S_{\min}(t) \cup j \in S_{\min}(t)}} p_{ij} - \lambda.$$

We conclude that $g'(\tilde{Q}(t)) \leq -h$, and hence, $g(\tilde{Q}(t)) \leq H_{\tilde{\mathbb{X}}}(t)$ for all $t \geq 0$. In particular, if $g(\tilde{Q}(0)) = 0$ and $|\tilde{\mathbb{X}}(0)| \leq 1$, it follows from the definition of $H_{\tilde{\mathbb{X}}}(\cdot)$ that $g(\tilde{Q}(t)) = 0$ for all $t \geq 0$.

The first statement of the lemma follows since for every hydrodynamic model solution $\tilde{\mathbb{X}}$ satisfying $|\tilde{\mathbb{X}}(0)| \leq 1$, it holds that $g(\tilde{Q}(0)) \leq 2/\min_{1 \leq j \leq S} p_j$. Hence, $H_{\tilde{\mathbb{X}}}(\cdot) \leq H(\cdot)$ for every hydrodynamic model solution $\tilde{\mathbb{X}}$ satisfying $|\tilde{\mathbb{X}}(0)| \leq 1$. □

This result implies that the hydrodynamically scaled queue length almost satisfies this property as well. The next result is an immediate consequence of Corollary 3.

Corollary 4 *Fix $\epsilon > 0$, $M > 0$ and $T > 0$. Then, for every $\omega \in \mathcal{K}^r$,*

$$g(Q^{r,m}(t)) \leq H(t) + \epsilon$$

for all $t \in [0, M]$ and $m < \sqrt{r}T$, where $H(\cdot)$ is as in Lemma 4. Moreover, if $g(\hat{Q}(0)) \to 0$ in probability as $r \to \infty$, then for all $\omega \in \mathcal{L}^r$ with

$$\mathcal{L}^r = \mathcal{K}^r \cap \left\{ \left| g(Q^{r,0}(0)) \right| \leq \epsilon \right\}$$

it holds that

$$\| g(Q^{r,0}(t)) \|_M \leq \epsilon,$$

and

$$\lim_{r \to \infty} \mathbb{P}(\mathcal{L}^r) = 1.$$

B.3 Multiplicative state-space collapse

The goal of this section is to show multiplicative state-space collapse for the SSC function defined in (75). To do so, we first need to relate the hydrodynamic and diffusion scaling. That is, we observe that

$$Q_j^{r,m}(t) = p_j \sqrt{\frac{\lambda r/\mu}{x_{r,m}}} \hat{Q}_j^r \left(\frac{m}{\sqrt{r}} + \frac{\sqrt{x_{r,m}}t}{r} \right) = \frac{p_j \sqrt{\lambda/\mu}}{y_{r,m}} \hat{Q}_j^r \left(\frac{1}{\sqrt{r}}(m + y_{r,m}t) \right),$$

where

$$y_{r,m} = \sqrt{\frac{x_{r,m}}{r}} = \max \left\{ \left| p \sqrt{\frac{\lambda}{\mu}} \hat{Q}^r \left(\frac{m}{\sqrt{r}} \right) \right|, \left| \hat{L}^r \left(\frac{m}{\sqrt{r}} \right) \right|, 1 \right\}$$

with

$$\hat{L}^r(t) = \frac{L^r(t) - r}{\sqrt{r}}.$$

Corollary 4 can be translated to the diffusion scaled process. Consider the SSC function $\hat{g} : \mathbb{R}^S \to \mathbb{R}$ defined as

$$\hat{g}(q) = \max_{1 \leq j \leq S} q_j - \min_{1 \leq j \leq S} q_j$$

with $q = (q_1, \ldots, q_S)$.

Corollary 5 *Fix $\epsilon > 0$, $M > 0$ and $T > 0$. Then for r large enough, and $\omega \in \mathcal{K}^r$,*

$$\hat{g}(\hat{Q}^r(t)) \leq \frac{y_{r,m}}{\sqrt{\lambda/\mu}} H\left(\frac{1}{y_{r,m}}(\sqrt{r}t - m)\right) + \epsilon \frac{y_{r,m}}{\sqrt{\lambda/\mu}}$$

for all $t \in [0, T]$ with $m \in \mathbb{N}$ such that

$$\frac{m}{\sqrt{r}} \leq t \leq \frac{m + y_{r,m}M}{\sqrt{r}}.$$

Also, for all $\omega \in \mathcal{L}^r$,

$$\|\hat{g}(\hat{Q}^r(t)\|_{My_{r,0}/\sqrt{r}} \leq \epsilon \frac{y_{r,0}}{\sqrt{\lambda/\mu}}.$$

Since $H(\cdot)$ is given as in (76), we observe that $H(t) = 0$ for all $t \geq 2/(h \min_{1\leq j\leq S} p_j)$. We would like to show that $(\sqrt{r}t - m)/y_{r,m}$ can be chosen large enough to obtain a very small upper bound, and use that property to show multiplicative state-space collapse.

Lemma 5 *Suppose $M \geq 2(N + 2)$ is fixed, and let*

$$m_r(t) = \min\left\{m \in \mathbb{N} : \frac{m}{\sqrt{r}} \leq t \leq \frac{m + y_{r,m}M}{\sqrt{r}}\right\}.$$

Then, for r large enough,

$$\frac{\sqrt{r}t - m_r(t)}{y_{r,m_r(t)}} \geq \frac{M}{2(N + 2)}$$

for every $\omega \in \mathcal{K}^r$ and $t \in (My_{r,0}/\sqrt{r}, T]$.

Proof For every $\omega \in \mathcal{K}^r$, by the definition of the set,

$$|\mathbb{X}^{r,m}(t_2) - \mathbb{X}^{r,m}(t_2)| \leq N|t_2 - t_1| + \epsilon,$$

for $t_1, t_2 \in [0, M]$ and $m < \sqrt{r}T$. In particular, for $t_2 = 1/y_{r,m}$, $t_1 = 0$ and $\epsilon \leq 1$,

$$\max\left\{\left|Q^r\left(\frac{m+1}{\sqrt{r}}\right) - Q^r\left(\frac{m}{\sqrt{r}}\right)\right|, \left|L^r\left(\frac{m+1}{\sqrt{r}}\right) - L^r\left(\frac{m}{\sqrt{r}}\right)\right|\right\} \leq \sqrt{x_{r,m}}\frac{N}{y_{r,m}} + \sqrt{x_{r,m}}.$$

Applying the reverse triangle inequality, we observe

$$\left|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r\left(\frac{m+1}{\sqrt{r}}\right)\right| - \left|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r\left(\frac{m}{\sqrt{r}}\right)\right| \leq \left|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r\left(\frac{m+1}{\sqrt{r}}\right) - p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r\left(\frac{m}{\sqrt{r}}\right)\right|$$
$$\leq N + y_{r,m} \leq (N + 1)y_{r,m},$$

and similarly,

$$\left|\hat{L}^r\left(\frac{m+1}{\sqrt{r}}\right)\right| - \left|\hat{L}^r\left(\frac{m}{\sqrt{r}}\right)\right| \leq (N+1)y_{r,m}.$$

Therefore, it always holds that

$$y_{r,m+1} \leq y_{r,m} + (N+1)y_{r,m} = (N+2)y_{r,m}.$$

For every $t \in (My_{r,0}/\sqrt{r}, T]$, it follows by the definition of $m_r(t)$ that

$$\sqrt{r}t \geq m_r(t) - 1 + y_{r,m_r(t)-1}M.$$

In particular,

$$\frac{\sqrt{r}t - m_r(t)}{y_{r,m_r(t)}} \geq \frac{y_{r,m_r(t)-1}M - 1}{y_{r,m_r(t)}} \geq \frac{M}{N+2} - \frac{1}{y_{r,m_r(t)}} \geq \frac{M}{2(N+2)},$$

where the last inequality follows since $M \geq 2(N+2)$. □

Next, we show the main result of this section.

Theorem 8 *Suppose $\hat{Q}^r(0) \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$. For every $T > 0$, $\epsilon > 0$ and $M < \infty$ with*

$$M \geq \max\left\{\frac{4(N+2)}{h\min_{1\leq j\leq S} p_j}, 2(N+2), 1\right\},$$

it holds that

$$\mathbb{P}\left(\frac{\sup_{My_{r,0}/\sqrt{r}\leq t\leq T}\hat{g}(\hat{Q}^r(t))}{\max\{\|\hat{Q}^r(t)\|_T, 1\}} > \epsilon\right) \to 0, \tag{77}$$

as $r \to \infty$. If, in addition, $\hat{g}(\hat{Q}^r(0)) \to 0$ in probability as $r \to \infty$, then for every $T > 0$,

$$\frac{\|\hat{g}(\hat{Q}^r(t))\|_T}{\max\{\|\hat{Q}^r(t)\|_T, 1\}} \xrightarrow{\mathbb{P}} 0, \tag{78}$$

as $r \to \infty$.

Proof Fix $\eta > 0$ and note that by construction there exists a r_0 such that, for all $r > r_0$,

$$\mathbb{P}(\mathcal{K}^r) > 1 - \eta.$$

For every $\omega \in \mathcal{K}^r$, we have derived bounds that only require that M is bounded. We note that M as in the statement of the theorem allows, for every $t \in [0, T]$,

$m/\sqrt{r} \le t \le (m + y_{r,m}M)/\sqrt{r}$ for some $m < \sqrt{r}M$. Moreover, it follows from Lemma 5 and (76) that

$$H\left(\frac{1}{y_{r,m_r(t)}}(\sqrt{r}t - m_r(t))\right) = 0$$

for all $t \in (My_{r,0}/\sqrt{r}, T]$. In view of Corollary 5, we obtain, for every $\epsilon > 0$,

$$\hat{g}(\hat{Q}^r(t)) \le \epsilon \frac{y_{r,m_r(t)}}{\sqrt{\lambda/\mu}}$$

for all $t \in (My_{r,0}/\sqrt{r}, T]$. Since, for all $t \in [0, T]$,

$$\begin{aligned} y_{r,m_r(t)} &= \max\left\{\left|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r\left(\frac{m_r(t)}{\sqrt{r}}\right)\right|, \left|\hat{L}^r\left(\frac{m_r(t)}{\sqrt{r}}\right)\right|, 1\right\} \\ &\le \max\left\{\left\|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r(t)\right\|_T, \left\|\hat{L}^r(t)\right\|_T, 1\right\}, \end{aligned}$$

we obtain, for every $\omega \in \mathcal{K}^r$,

$$\frac{\sup_{\frac{My_{r,0}}{\sqrt{r}} \le t \le T} \hat{g}(\hat{Q}^r(t))}{\max\left\{\left\|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r(t)\right\|_T, \left\|\hat{L}^r(t)\right\|_T, 1\right\}} \le \frac{\epsilon}{\sqrt{\lambda/\mu}}.$$

Note that, for every $t \ge 0$,

$$|\hat{L}^r(t)| = \left|\sum_{j=1}^{S}\left(p_j\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r_j(t) - \beta\sqrt{\frac{\lambda}{\mu}}\right)^+\right| \le \sqrt{\frac{\lambda}{\mu}}\left(|\hat{Q}^r(t)| + S|\beta|\right). \tag{79}$$

Moreover, since $\epsilon > 0$ is arbitrary, we can conclude that (77) holds.

If $|\hat{Q}^r(0)| \xrightarrow{\mathbb{P}} 0$, it follows from Corollary 5 that, for all $\omega \in \mathcal{L}^r$ and $t \in [0, My_{r,0}/\sqrt{r}]$,

$$\hat{g}(\hat{Q}^r(t)) \le \frac{\epsilon}{\sqrt{\lambda/\mu}} y_{r,0} \le \frac{\epsilon}{\sqrt{\lambda/\mu}} \max\left\{\left\|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r(t)\right\|_T, \left\|\hat{L}^r(t)\right\|_T, 1\right\}.$$

Since $\epsilon > 0$ is arbitrary, together with (77) and (79), we obtain (78). □

Remark 5 Note that the bounds in Theorem 8 are obtained for every fixed $T > 0$. Yet, from the proof it is clear that one has the following slightly more general result:

Suppose $\hat{Q}^r(0) \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$. For every $\epsilon > 0$, $M < \infty$ as in Theorem 8, and $t_r \in (My_{r,0}/\sqrt{r}, \infty)$,

$$\mathbb{P}\left(\frac{\sup_{My_{r,0}/\sqrt{r} \le s \le t_r} \hat{g}(\hat{Q}^r(s))}{\max\{\|\hat{Q}^r(s)\|_{t_r}, 1\}} > \epsilon\right) \to 0, \tag{80}$$

as $r \to \infty$. If, in addition, $\hat{g}(\hat{Q}^r(0)) \to 0$ in probability as $r \to \infty$, then for every $t_r \in (My_{r,0}/\sqrt{r}, \infty)$,

$$\frac{\|\hat{g}(\hat{Q}^r(s))\|_{t_r}}{\max\{\|\hat{Q}^r(t)\|_{t_r}, 1\}} \xrightarrow{\mathbb{P}} 0, \tag{81}$$

as $r \to \infty$. In other words, the interval over which the state-space collapse is considered can also be chosen as a sequence of intervals indexed by r.

B.4 Strong state-space collapse

Although Theorem 8 shows multiplicative state-space collapse, our interest lies in the strong state-space collapse as is stated in Theorem 6. To do so, it suffices to show that the denominators in Theorem 8 are bounded in a probabilistic sense. More specifically, $\|\hat{Q}^r(t)\|_T$ should satisfy the compact containment property. Before doing so, we prove a result that shows that even if the diffusion-scaled queue lengths are initially not necessarily close to one another, these queue lengths do not explode for a sufficiently short period of time.

Lemma 6 *Suppose $\hat{Q}^r(0) \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$, and $M \in [0, \infty)$. Then,*

$$\lim_{K\to\infty} \lim_{r\to\infty} \mathbb{P}\left(\|\hat{Q}^r(t)\|_{My_{r,0}/\sqrt{r}} > K\right) = 0.$$

Proof Fix $\epsilon \in (0, 1)$ small. First, note that

$$\begin{aligned}&\mathbb{P}\left(\|\hat{Q}^r(t)\|_{My_{r,0}/\sqrt{r}} > K\right)\\ &\quad\le \mathbb{P}\left(\|\hat{Q}^r(t)\|_{My_{r,0}/\sqrt{r}} > K; \max\left\{|\hat{Q}^r(0)|, y_{r,0}\right\} \le \epsilon K\right) + \mathbb{P}\left(\max\left\{|\hat{Q}^r(0)|, y_{r,0}\right\} > \epsilon K\right).\end{aligned} \tag{82}$$

By definition,

$$y_{r,0} = \max\left\{\left|p\sqrt{\frac{\lambda}{\mu}}\hat{Q}^r(0)\right|, \left|\hat{L}^r(0)\right|, 1\right\}.$$

Since $\hat{Q}^r(0) \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$,

$$\lim_{K\to\infty} \lim_{r\to\infty} \mathbb{P}\left(|\hat{Q}^r(0)| > \epsilon K\right) = 0,$$

and due to the definition of $y_{r,0}$ and (79) for $t = 0$, this implies

$$\lim_{K\to\infty} \lim_{r\to\infty} \mathbb{P}\left(\max\left\{|\hat{Q}^r(0)|, y_{r,0}\right\} > \epsilon K\right) = 0.$$

To bound the first term in (82) as well, we observe that the queue length at some time is trivially bounded by

$$|Q^r(t)| \le |Q^r(0)| + \max\left\{\sum_{\{i,j\}\in E} A_{ij}(rt), \max_{1\le j\le S}\{S_j(F_j^r t)\}\right\}.$$

We observe that $F_j^r \le (1+\epsilon)\lambda r/\mu$ for r large enough. Moreover, if $\{A(t), t \ge 0\}$ denotes a Poisson process with rate λ, then due to the properties of the Poisson process it holds that $\sum_{\{i,j\}\in E} A_{ij}(\cdot) \stackrel{d}{=} A(\cdot)$. In terms of the diffusion scaling, the above bound yields, for all $t \ge 0$,

$$|\hat{Q}^r(t)| \le |\hat{Q}^r(0)| + \frac{\max\left\{A(rt), \max_{1\le j\le S}\{S_j((1+\epsilon)\lambda r/\mu t)\}\right\}}{\min_{1\le j\le S} p_j\sqrt{\lambda r/\mu}}.$$

Therefore, using this bound for $t = My_{r,0}/\sqrt{r} \le \epsilon MK/\sqrt{r}$ and noting that Poisson processes are (non-decreasing) counting processes,

$$\begin{aligned}&\mathbb{P}\left(\|\hat{Q}^r(t)\|_{My_{r,0}/\sqrt{r}} > K; \max\left\{|\hat{Q}^r(0)|, y_{r,0}\right\} \le \epsilon K\right)\\ &\quad\le \mathbb{P}\left(\frac{\max\left\{A(\epsilon MK\sqrt{r}), \max_{1\le j\le S}\left\{S_j\left(\epsilon(1+\epsilon)\lambda/\mu MK\sqrt{r}\right)\right\}\right\}}{\min_{1\le j\le S} p_j\sqrt{\lambda r/\mu}} > (1-\epsilon)K\right).\end{aligned}$$

Due to the LLN, we observe that

$$\lim_{K\to\infty} \lim_{r\to\infty} \mathbb{P}\left(\frac{A(\epsilon MK\sqrt{r})}{\min_{1\le j\le S} p_j\sqrt{\lambda/\mu}K\sqrt{r}} > (1-\epsilon)\right) = 0$$

for $\epsilon > 0$ small enough (for example, for $\epsilon < 1 - M/(M + \sqrt{\lambda/\mu}\min_{1\le j\le S} p_j)$). Similarly, due to the LLN,

$$\lim_{K\to\infty} \lim_{r\to\infty} \sum_{j=1}^{S} \mathbb{P}\left(\frac{S_j\left(\epsilon(1+\epsilon)\lambda/\mu MK\sqrt{r}\right)}{\min_{1\le j\le S} p_j\sqrt{\lambda/\mu}K\sqrt{r}} > (1-\epsilon)\right) = 0$$

for $\epsilon > 0$ small enough (for example, for $\epsilon < \min_{1\le j\le S} p_j/(M\sqrt{\lambda/\mu} + \min_{1\le j\le S} p_j)$). We conclude that the first term in (82) converges to zero, i.e.,

$$\lim_{K\to\infty}\lim_{r\to\infty}\mathbb{P}\left(\|\hat{Q}^r(t)\|_{My_{r,0}/\sqrt{r}} > K; \max\left\{|\hat{Q}^r(0)|, y_{r,0}\right\} \le \epsilon K\right) = 0,$$

and hence, the result follows. □

Next, we show that the process $\hat{Q}^r(\cdot)$ satisfies the compact containment property.

Proposition 4 *Suppose $|\hat{Q}^r(0)| \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$. Then, for every $T > 0$ and $\epsilon > 0$,*

$$\lim_{K\to\infty}\lim_{r\to\infty}\mathbb{P}\left(\|\hat{Q}^r(t)\|_T > K\right) = 0. \tag{83}$$

Proof Fix $\epsilon \in (0, 1/3)$ small, and let $K > \max\{2|\beta| + 2, 2|\gamma| + 2\}$. Introduce the sequence of stopping times

$$\hat{\tau}_K^r = \inf\left\{t \ge 0 : \max_{1\le j\le S}\bar{Q}_j^r(t) > K\right\}, \quad \hat{T}_K^r = \sup\left\{0 \le t \le \hat{\tau}_k : \min_{1\le j\le S}\bar{Q}_j^r(t) \le K/2\right\},$$

and similarly,

$$\check{\tau}_K^r = \inf\left\{t \ge 0 : \min_{1\le j\le S}\bar{Q}_j^r(t) < -K\right\}, \quad \check{T}_K^r = \sup\left\{0 \le t \le \check{\tau}_k : \max_{1\le j\le S}\bar{Q}_j^r(t) \ge -K/2\right\}.$$

Clearly,

$$\mathbb{P}\left(\|\hat{Q}^r(t)\|_T > K\right) \le \mathbb{P}\left(\hat{\tau}_K^r \le \check{\tau}_K^r \le T\right) + \mathbb{P}\left(\check{\tau}_K^r \le \hat{\tau}_K^r \le T\right). \tag{84}$$

In order to improve the readability of the proof, we first present a proof in the case when $\hat{g}(\hat{Q}^r(0)) \to 0$ in probability as $r \to \infty$. We then comment on the changes needed to adapt the proof to the case when this condition does not necessarily hold.

Case I: $\hat{g}(\hat{Q}^r(0)) \to 0$ in probability as $r \to \infty$.

Since $|\hat{Q}^r(0)| \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$, it holds that

$$\lim_{K\to\infty}\lim_{r\to\infty}\mathbb{P}\left(|\hat{Q}^r(0)| > K/2\right) = 0,$$

and hence, we can assume that both $\hat{\tau}_K^r > \hat{T}_K^r > 0$ and $\check{\tau}_K^r > \check{T}_K^r > 0$ (for K large enough). Moreover, for every $t \le \min\{\check{\tau}_K^r, \hat{\tau}_K^r\}$,

$$\frac{\|\hat{g}(\hat{Q}^r(t))\|_{\min\{\check{\tau}_K^r,\hat{\tau}_K^r\}}}{\max\{\|\hat{Q}^r(t)\|_{\min\{\check{\tau}_K^r,\hat{\tau}_K^r\}}, 1\}} \le \epsilon \quad \Rightarrow \quad \max_{1\le i\le S}\hat{Q}_i^r(t) - \min_{1\le i\le S}\hat{Q}_i^r(t) \le \epsilon K. \tag{85}$$

Next, we provide bounds for the two ways of crossing the boundary K separately. First, we consider the first term in (84). We observe

$$\begin{aligned}&\mathbb{P}\left(\hat{\tau}_K^r \le \check{\tau}_K^r \le T\right)\\ &\quad\le \mathbb{P}\left(\hat{\tau}_K^r \le \check{\tau}_K^r \le T\,;\, \frac{\|\hat{g}(\hat{Q}^r(t))\|_{\hat{\tau}_K^r}}{\max\{\|\hat{Q}^r(t)\|_{\hat{\tau}_K^r}, 1\}} \le \epsilon\right) + \mathbb{P}\left(\frac{\|\hat{g}(\hat{Q}^r(t))\|_{\hat{\tau}_K^r}}{\max\{\|\hat{Q}^r(t)\|_{\hat{\tau}_K^r}, 1\}} > \epsilon\right).\end{aligned}$$

Due to Theorem 8 and (81),

$$\lim_{r\to\infty} \mathbb{P}\left(\frac{\|\hat{g}(\hat{Q}^r(t))\|_{\hat{\tau}_K^r}}{\max\{\|\hat{Q}^r(t)\|_{\hat{\tau}_K^r}, 1\}} > \epsilon\right) = 0.$$

To bound the first term, define the process $\{\hat{Q}_\Sigma^r(t), t \ge 0\}$ with

$$\hat{Q}_\Sigma^r(t) = \frac{\sum_{j=1}^S \left(Q_j^r(t) - p_j\lambda r/\mu\right)}{\sqrt{\lambda r/\mu}} = \sum_{j=1}^S p_j \hat{Q}_j^r(t).$$

We observe that

$$\min_{1\le j\le S} \hat{Q}_j^r(t) \le \hat{Q}_\Sigma^r(t) \le \max_{1\le j\le S} \hat{Q}_j^r(t). \tag{86}$$

Due to the system identities, we observe that, for every $t \in [\hat{T}_K^r, \hat{\tau}_K^r]$,

$$\hat{Q}_\Sigma^r(t) = \hat{Q}_\Sigma^r(\hat{T}_K^r) + \frac{\sum_{\{i,j\}\in E} A_{ij}^r(t) - A_{ij}^r(\hat{T}_K^r)}{\sqrt{\lambda r/\mu}} - \frac{\sum_{j=1}^S D_j^r(t) - D_j^r(\hat{T}_K^r)}{\sqrt{\lambda r/\mu}}.$$

We note that, due to the properties of the Poisson process,

$$\sum_{\{i,j\}\in E} A_{ij}^r(t) - A_{ij}^r(\hat{T}_K^r) \le_{\mathrm{ST}} \sum_{\{i,j\}\in E} \Lambda_{ij}^r(rt) - \Lambda_{ij}^r(r\hat{T}_K^r) \stackrel{d}{=} \Lambda(rt) - \Lambda(r\hat{T}_K^r)$$

with $\{\Lambda(t), t \ge 0\}$ an (independent) Poisson process with rate λ. Moreover, since for all $t \in [\hat{T}_K^r, \hat{\tau}_K^r]$ it holds that $\hat{Q}_j^r(t) \ge \gamma$ for every $i \in \{1, \ldots, S\}$,

$$\sum_{j=1}^S D_j^r(t) - D_j^r(\hat{T}_K^r) = \sum_{j=1}^S S_j(F_j^r t) - S_j(F_j^r \hat{T}_K^r).$$

Using the FCLT, we observe that

$$\frac{\Lambda(rt) - \Lambda(r\hat{T}_K^r) - \sum_{j=1}^S S_j(F_j^r t) - S_j(F_j^r \hat{T}_K^r)}{\sqrt{\lambda r/\mu}} \xrightarrow{d} \mathrm{BM}(t) - \mathrm{BM}(\hat{T}_K^r) - \gamma\mu(t - \hat{T}_K^r)$$

as $r \to \infty$, where $\{\mathrm{BM}(t), t \geq 0\}$ is a Brownian motion with zero mean and finite variance (independent of K). Finally, by the definitions of the stopping times, and in view of (85) and (86), for all $t \in [\hat{T}_K^r, \hat{\tau}_K^r]$,

$$\hat{Q}_\Sigma^r(\hat{\tau}_K^r) - \hat{Q}_\Sigma^r(\hat{T}_K^r) \geq (1-\epsilon)K - (1+\epsilon)K/2 = (1-3\epsilon)K/2.$$

We conclude that

$$\begin{aligned}
&\lim_{r\to\infty} \mathbb{P}\left(\hat{\tau}_K^r \leq \check{\tau}_K^r \leq T \,;\, \frac{\|\hat{g}(\hat{Q}^r(t))\|_{\hat{\tau}_K^r}}{\max\{\|\hat{Q}^r(t)\|_{\hat{\tau}_K^r}, 1\}} \leq \epsilon \right) \\
&\quad \leq \mathbb{P}\left(\sup_{0\leq s\leq t\leq T} \mathrm{BM}(t) - \mathrm{BM}(s) - \gamma\mu(t-s) \geq \frac{1-3\epsilon}{2} K \right),
\end{aligned}$$

which converges to zero as $K \to \infty$ since $\epsilon \in (0, 1/3)$.

The analysis of the second term in (84) uses similar arguments as the first term. We observe

$$\begin{aligned}
&\mathbb{P}\left(\check{\tau}_K^r \leq \hat{\tau}_K^r \leq T\right) \\
&\quad \leq \mathbb{P}\left(\check{\tau}_K^r \leq \hat{\tau}_K^r \leq T \,;\, \frac{\|\hat{g}(\hat{Q}^r(t))\|_{\check{\tau}_K^r}}{\max\{\|\hat{Q}^r(t)\|_{\check{\tau}_K^r}, 1\}} \leq \epsilon \right) + \mathbb{P}\left(\frac{\|\hat{g}(\hat{Q}^r(t))\|_{\check{\tau}_K^r}}{\max\{\|\hat{Q}^r(t)\|_{\check{\tau}_K^r}, 1\}} > \epsilon \right).
\end{aligned}$$

Again, due to Theorem 8 and (81),

$$\lim_{r\to\infty} \mathbb{P}\left(\frac{\|\check{g}(\hat{Q}^r(t))\|_{\check{\tau}_K^r}}{\max\{\|\check{Q}^r(t)\|_{\check{\tau}_K^r}, 1\}} > \epsilon \right) = 0.$$

Due to the system identities, we observe that, for every $t \in [\check{T}_K^r, \check{\tau}_K^r]$,

$$\hat{Q}_\Sigma^r(t) = \hat{Q}_\Sigma^r(\check{T}_K^r) + \frac{\sum_{\{i,j\}\in E} A_{ij}^r(t) - A_{ij}^r(\check{T}_K^r)}{\sqrt{\lambda r/\mu}} - \frac{\sum_{j=1}^S D_j^r(t) - D_j^r(\check{T}_K^r)}{\sqrt{\lambda r/\mu}}.$$

Due to the definitions of the stopping times, we observe that, for all $t \in [\check{T}_K^r, \check{\tau}_K^r]$, it holds that $\hat{Q}_j^r(t) \leq \beta$ for every $j \in \{1, \ldots, S\}$, and hence, $L^r(t) = r$. Therefore, due to the properties of the Poisson process,

$$\sum_{\{i,j\}\in E} A_{ij}^r(t) - A_{ij}^r(\check{T}_K^r) = \sum_{\{i,j\}\in E} \Lambda_{ij}^r(rt) - \Lambda_{ij}^r(r\check{T}_K^r) \stackrel{d}{=} \Lambda(rt) - \Lambda(r\check{T}_K^r)$$

with $\{\Lambda(t), t \geq 0\}$ a Poisson process with rate λ. Moreover,

$$\sum_{j=1}^S D_j^r(t) - D_j^r(\check{T}_K^r) \leq_{\mathrm{ST}} \sum_{j=1}^S S_j(F_j^r t) - S_j(F_j^r \check{T}_K^r).$$

Using the FCLT, we observe again that

$$\frac{A(rt) - A(r\breve{T}_K^r) - \sum_{j=1}^{S} S_j(F_j^r t) - S_j(F_j^r \breve{T}_K^r)}{\sqrt{\lambda r/\mu}} \xrightarrow{d} \mathrm{BM}(t) - \mathrm{BM}(\breve{T}_K^r) - \gamma\mu(t - \breve{T}_K^r)$$

as $r \to \infty$, where we recall that $\{\mathrm{BM}(t), t \geq 0\}$ is a Brownian motion with zero mean and finite variance (independent of K). Finally, by the definition of the stopping times, and in view of (85) and (86), it holds for all $t \in [\breve{T}_K^r, \breve{\tau}_K^r]$ that

$$\hat{Q}_\Sigma^r(\breve{\tau}_K^r) - \hat{Q}_\Sigma^r(\breve{T}_K^r) \leq -(1-\epsilon)K - (-(1+\epsilon)K/2) = -\frac{1-3\epsilon}{2}K.$$

We conclude that

$$\lim_{r\to\infty} \mathbb{P}\left(\breve{\tau}_K^r \leq \hat{\tau}_K^r \leq T\ ;\ \frac{\|\hat{g}(\hat{Q}^r(t))\|_{\breve{\tau}_K^r}}{\max\{\|\hat{Q}^r(t)\|_{\breve{\tau}_K^r}, 1\}} \leq \epsilon\right)$$
$$\leq \mathbb{P}\left(\sup_{0\leq s\leq t\leq T} \mathrm{BM}(t) - \mathrm{BM}(s) - \gamma\mu(t-s) \leq -\frac{1-3\epsilon}{2}K\right),$$

which also converges to zero as $K \to \infty$ since $\epsilon \in (0, 1/3)$. Since this holds for both of the two summed probabilities in Theorem (84), we conclude that (83) holds.

Case II: general case, i.e., when we do not assume that $\hat{g}(\hat{Q}^r(0)) \to 0$ in probability as $r \to \infty$.

Let $M \in [1, \infty)$ be fixed and satisfy the property as in Theorem 8. Since $|\hat{Q}^r(0)| \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$ and due to Lemma 6, it holds that

$$\lim_{K\to\infty} \lim_{r\to\infty} \mathbb{P}\left(\|\hat{Q}^r(t)\|_{My_{r,0}/\sqrt{r}} > K/2\right) = 0.$$

Therefore, we can assume that both $\hat{\tau}_K^r > \hat{T}_K^r > My_{r,0}/\sqrt{r}$ and $\breve{\tau}_K^r > \breve{T}_K^r > My_{r,0}/\sqrt{r}$ (for K large enough). The proof in this general case is then completely analogous to that in the previous one: $\|\hat{g}(\hat{Q}^r(t))\|_{\breve{\tau}_K^r}$ needs to be replaced with $\sup_{t\in(My_{r,0}/\sqrt{r},\breve{\tau}_K^r]}|\hat{g}(\hat{Q}^r(t))|$, and $\|\hat{g}(\hat{Q}^r(t))\|_{\hat{\tau}_K^r}$ with $\sup_{t\in(My_{r,0}/\sqrt{r},\hat{\tau}_K^r]}|\hat{g}(\hat{Q}^r(t))|$. □

Next, we prove our main result stated as in Theorem 6.

Proof of Theorem 6 Equation (34) is a consequence of Theorem 8 and Proposition 4. To prove (33), note that, for every $\epsilon > 0$ and any sequence $\{K^r, r \in \mathbb{N}\}$,

$$\mathbb{P}\left(\sup_{K^r/\sqrt{r}\leq t\leq T} \hat{g}(\hat{Q}^r(t)) > \epsilon\right)$$
$$= \mathbb{P}\left(\sup_{K^r/\sqrt{r}\leq t\leq T} \hat{g}(\hat{Q}^r(t)) > \epsilon\ ;\ K^r > My_{r,0}\right) + \mathbb{P}\left(K^r \leq My_{r,0}\right)$$

$$\leq \mathbb{P}\left(\sup_{My_{r,0}/\sqrt{r}\leq t\leq T} \hat{g}(\hat{Q}^r(t)) > \epsilon\right) + \mathbb{P}\left(K^r \leq My_{r,0}\right).$$

Theorem 8 and Proposition 4 imply that, for every $\epsilon > 0$,

$$\lim_{r\to\infty} \mathbb{P}\left(\sup_{My_{r,0}/\sqrt{r}\leq t\leq T} \hat{g}(\hat{Q}^r(t)) > \epsilon\right) = 0,$$

Moreover, for any sequence $\{K^r, r \in \mathbb{N}\}$ for which $K^r = o(\sqrt{r})$ with $K^r \to \infty$ as $r \to \infty$, it holds that

$$\lim_{r\to\infty} \mathbb{P}\left(K^r \leq My_{r,0}\right) = 0.$$

by the definition of $y_{r,0}$, (79) and since $\hat{Q}^r(0) \to \hat{Q}(0)$ for some random vector $\hat{Q}(0)$. We conclude that (33) holds as well. □

References

1. Ata, B., Kumar, S.: Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies. Ann. Appl. Probab. **15**(1A), 331–391 (2005)
2. Bienstock, D.: Electrical Transmission System Cascades and Vulnerability. Society for Industrial and Applied Mathematics, Philadelphia (2015)
3. Billingsley, P.: Convergence of Probability Measures, Wiley Series in Probability and Statistics: Probability and Statistics, 2nd edn. Wiley, New York (1999)
4. Borst, S.C., Mandelbaum, A., Reiman, M.I.: Dimensioning large call centers. Oper. Res. **52**(1), 17–34 (2004)
5. Bramson, M.: State space collapse with application to heavy traffic limits for multiclass queueing networks. Queueing Syst. Theory Appl. **30**(1–2), 89–148 (1998)
6. Bramson, M.: State space collapse with application to heavy traffic limits for multiclass queueing networks. Queueing Syst. **30**(1), 89–140 (1998)
7. Browne, S. and Whitt, W.: Piecewise-linear diffusion processes. In: J. Dshalalow (ed.), Advances in Queueing, pp. 463–480. CRC Press, Boca Raton, FL (1995)
8. Dai, J.G.: Stability of fluid and stochastic processing networks. MaPhySto Miscellanea Publication, No. 9 (1999)
9. Dai, J.G., Tezcan, T.: State space collapse in many-server diffusion limits of parallel server systems. Math. Oper. Res. **36**(2), 271–320 (2011)
10. de Véricourt, F., Jennings, O.: Dimensioning large-scale membership services. Oper. Res. **55**(1), 173–187 (2008)
11. de Véricourt, F., Jennings, O.: Nurse staffing in medical units: a queueing perspective. Oper. Res. **59**(6), 1320–1331 (2011)
12. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**, 567–588 (1981)
13. Khudyakov, P., Feigin, P.D., Mandelbaum, A.: Designing a call center with an IVR (interactive voice response). Queueing Syst. **66**(3), 215–237 (2010)
14. National Academies of Sciences, E., Medicine: Analytic Research Foundations for the Next-Generation Electric Grid. The National Academies Press, Washington, DC (2016)
15. Sloothaak, F., Cruise, J.R., Shneer, V., Vlasiou, M., Zwart, B.: Complete resource pooling of a load balancing policy for a network of battery swapping stations. Preprint arXiv:1902.04392 (2019)
16. Sun, B., Sun, X., Tsang, D.H.K., Whitt, W.: Optimal battery purchasing and charging strategy at electric vehicle battery swap stations. Eur. J. Oper. Res. **279**(2), 524–539 (2019)

17. Sun, B., Tan, X., Tsang, D.H.K.: Optimal charging operation of battery swapping stations with QoS guarantee. In: 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 13–18 (2014)
18. Sun, B., Tan, X., Tsang, D.H.K.: Optimal charging operation of battery swapping and charging stations with QoS guarantee. IEEE Trans. Smart Grid **9**(5), 4689–4701 (2018)
19. Tan, X., Sun, B., Tsang, D.H.K.: Queueing network models for electric vehicle charging station with battery swapping. In: 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 1–6 (2014)
20. Tan, X., Sun, B., Wu, Y., Tsang, D.H.K.: Asymptotic performance evaluation of battery swapping and charging station for electric vehicles. Perform. Eval. **119**, 43–57 (2018)
21. Tezcan, T.: Optimal control of distributed parallel server systems under the Halfin and Whitt regime. Math. Oper. Res. **33**(1), 51–90 (2008)
22. van der Boor, M., Borst, S.C., van Leeuwaarden, J.S.H., Mukherjee, D.: Scalable load balancing in networked systems: a survey of recent advances. Preprint arXiv:1806.05444 (2018)
23. van Leeuwaarden, J.S.H., Mathijsen, B.W.J., Sloothaak, F., Yom-Tov, G.B.: The restricted Erlang-R queue: finite-size effects in service systems with returning customers. Preprint arXiv:1612.07088 (2016)
24. Williams, R.J.: Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. Queueing Syst. Theory Appl. **30**(1–2), 27–88 (1998)
25. Yom-Tov, G.B., Mandelbaum, A.: Erlang-R: a time-varying queue with reentrant customers, in support of healthcare staffing. Manuf. Serv. Oper. Manag. **16**(2), 283–299 (2014)
26. Zhang, B., van Leeuwaarden, J.S.H., Zwart, B.: Staffing call centers with impatient customers: refinements to many-server asymptotics. Oper. Res. **60**(2), 461–474 (2012)