The Annals of Applied Probability
2019, Vol. 29, No. 4, 2114–2174
https://doi.org/10.1214/18-AAP1444

THE HYDRODYNAMIC LIMIT OF A RANDOMIZED LOAD BALANCING NETWORK

BY REZA AGHAJANI[1] AND KAVITA RAMANAN[2]

University of California, San Diego and Brown University

Randomized load balancing networks arise in a variety of applications, and allow for efficient sharing of resources, while being relatively easy to implement. We consider a network of parallel queues in which incoming jobs with independent and identically distributed service times are assigned to the shortest queue among a subset of d queues chosen uniformly at random, and leave the network on completion of service. Prior work on dynamical properties of this model has focused on the case of exponential service distributions. In this work, we analyze the more realistic case of general service distributions. We first introduce a novel particle representation of the state of the network, and characterize the state dynamics via a countable sequence of interacting stochastic measure-valued evolution equations. Under mild assumptions, we show that the sequence of scaled state processes converges, as the number of servers goes to infinity, to a hydrodynamic limit that is characterized as the unique solution to a countable system of coupled deterministic measure-valued equations. As a simple corollary, we also establish a propagation of chaos result that shows that finite collections of queues are asymptotically independent. The general framework developed here is potentially useful for analyzing a larger class of models arising in diverse fields including biology and materials science.

CONTENTS

Received October 2017; revised July 2018.

[1] Supported in part by NSF Grant CMMI-1538706.

[2] Supported in part by ARO Grant W911NF-12-1-0222.

MSC2010 subject classifications. 60K25, 60K35, 90B15, 90B22, 68M20, 60F17.

Key words and phrases. Randomized load balancing, hydrodynamic limit, fluid limit, power of two choices, stochastic network, measure-valued processes, interacting particle system, propagation of chaos.

1. Introduction.

1.1. *Background and motivation.* Randomized load balancing is a method for the efficient sharing of resources in networking systems that is relatively easy to implement, and used in a variety of applications such as, for example, hash tables in data switches, parallel computing [30] and wireless networks [19]. In this article, we introduce a mathematical framework for the analysis of a class of large-scale parallel server load balancing networks in the presence of general service times, with the specific goal of obtaining a tractable characterization of the hydrodynamic limit of randomized join-the-shortest-queue networks, as the number of servers goes to infinity. Past work on dynamical properties of this model has essentially been restricted to the case of exponential service distributions. A key component of our approach that allows us to handle general service distributions is a description of its dynamics via a sequence of interacting stochastic measure-valued evolution equations, which are amenable to analysis. Our framework can be generalized and we expect similar representations to also be useful for the study of other load-balancing models [29] as well as models arising in population biology and materials science.

In the randomized join-the-shortest-queue network model, also referred to as the supermarket model, jobs with independent and identically distributed (i.i.d.) service times arrive according to a renewal process with rate λN to a network of

N identical servers in parallel, each with an infinite-capacity queue. Upon arrival of a job, d queues are sampled independently and uniformly at random (with replacement) and the job is routed to the shortest queue among those sampled, with ties broken uniformly at random. Each server processes jobs from its queue in a first-come first-serve (FCFS) manner, a server never idles when there is a job in its queue and jobs leave the network on completion of service. The arrival process and service times are assumed to be mutually independent, and service times of jobs have finite mean which, without loss of generality, will be taken to be one. We refer to this model as the (N-server) SQ(d) model. A positive feature of this algorithm is that its implementation does not require much system memory.

Several results are known when the arrival process is Poisson with rate $\lambda < 1$ and the service time is exponential (with unit mean). When $d = 1$, the model reduces to a system of N independent single-server queues with exponential service times, for which it is a classical result that the stationary distribution of the length of a typical queue is geometric, and thus has an exponentially decaying tail. When $d = 2$, the stationary distribution of a typical queue is not exactly computable, but it was shown in [33] (also see [30] for the extension to $d > 2$) that as the number of servers goes to infinity, the limit of the stationary distributions of a typical queue has a doubly exponential tail. This shows that introducing just a little bit of random choice leads to a dramatic improvement in performance in equilibrium, a phenomenon that has been dubbed the "power of two choices" and has led to substantial interest in this class of randomized load balancing schemes.

The analysis in [33] proceeds by representing the dynamics of the N-server network by a Markov chain that keeps track of the fraction of queues that have ℓ or more jobs at time t, for each positive integer ℓ, and then applying the so-called ODE method (see Theorem 11.2.1 of [17]) to show that, as $N \to \infty$, the sequence of Markov chains converges weakly (on finite time intervals) to the unique solution of a countable system of coupled $[0, 1]$-valued ordinary differential equations (ODEs). Further tightness estimates are then used to prove convergence of the stationary distributions to the unique invariant state of the ODE. This basic approach was subsequently used to analyze various relevant modifications of the supermarket model [18, 23, 29]. Other theoretical results on the SQ(d) model with exponential service distributions in this asymptotic regime include [20, 28, 31].

However, measurements in different applications have shown that service times are typically not exponentially distributed [13, 15, 26, 27]. In this case, the ODE method is no longer directly applicable because in order to describe the future evolution of the system, it is not sufficient to keep track of the fraction of queues with ℓ jobs at any time. For each job in service, one has also to keep track of additional information such as its age (the amount of time the job has spent in service) or its residual service time. In a system with N servers, this requires keeping track of N additional nonnegative random variables, and thus the dimension of the Markovian state representation grows with N, which is not conducive to obtaining a limit theorem. Our goal is to develop a general framework for the analysis of this model and related models, that in particular enables an intuitive and tractable description of the hydrodynamic limit.

1.2. *Discussion of results.* To achieve our goal, we introduce a novel interacting particle representation of the state of the network that allows for a description of the dynamics of all N-server systems on a common (infinite-dimensional) state space. Specifically, we represent the state of an N-server SQ(d) network at any time t in terms of an infinite sequence of finite measures $\nu^{(N)}(t) = (\nu_\ell^{(N)}(t); \ell \in \mathbb{N})$, where $\nu_\ell^{(N)}(t)$ is the measure that has a "particle" or unit delta mass at the age of each job that is in service at a queue of length greater than or equal to ℓ at time t, where length denotes the number of jobs either waiting or in service. We then characterize the dynamics of the N-server SQ(d) model in terms of a coupled system of interacting stochastic measure-valued evolution equations (see Proposition 4.5). The main result of this article, Theorem 2.6, shows that under general conditions on the service time distribution and arrival processes, as $N \to \infty$, the sequence of scaled state processes $\nu^{(N)}/N$ converges weakly to the unique solution of a coupled system of deterministic measure-valued equations, which we refer to as the hydrodynamic equations (see Definition 2.1). While measure-valued representations keeping track of ages of jobs have recently been introduced to analyze certain many-server models such as the GI/GI/N model studied in [22, 24], to the best of our knowledge, this is the first work to consider a representation in terms of a countable sequence of *interacting* measure-valued stochastic processes. This poses several new technical challenges because the dynamics is significantly more complicated than in the models considered in [22, 24], in large part due to the state-dependent routing structure of the $SQ(d)$ network model and the interactions between the various component processes. In addition, the routing term contains components that evolve on different time-scales (as elaborated in Section 6.2.2), making its convergence analysis more challenging.

The tractable dynamical characterization of the limit via the hydrodynamic equations is one of the main contributions of this article. Our characterization provides qualitative insight into transient network performance, whose value is further illustrated in the related works [2–4], where, under some additional conditions on the service distribution, a reduced countable system of interacting classical PDEs is shown to capture essential performance characteristics of the limit system such as queue lengths and virtual waiting times. In [4], the numerical solution of the PDE is used to identify nonintuitive behavior of load-balancing networks, such as, for example, showing that under the SQ(d) routing algorithm, backlog in the system is cleared faster when the service distribution is heavy-tailed rather than light-tailed (with the same mean), and that for light-tailed service distributions, relaxation times (suitably defined) under the SQ(d) algorithm are significantly shorter than under random routing. Thus, our approach yields a PDE method for analyzing randomized load balancing networks, which generalizes the more classical ODE method that is valid only in the presence of exponential service distributions, and can be adapted to study several other models.

As a simple corollary of our main result, we also obtain a "propagation of chaos" result for finite time intervals, showing asymptotic independence of any finite set of servers at any given time as $N \to \infty$, when the initial conditions are exchangeable (see Corollary 2.8). A similar result has been obtained in Proposition 7.1 of [10] using completely different techniques (though only in the more restrictive setting of Poisson arrivals and i.i.d. fixed initial conditions). However, no dynamical characterization of the system was obtained in [10]. Indeed, just establishing propagation of chaos (over finite time intervals) is not sufficient to obtain a convenient characterization of the dynamics that can be used to provide qualitative insight into transient network performance such as our paper, along with [4], does.

Although the main focus of our paper is to understand transient behavior of the $SQ(d)$ load balancing network, we believe that this paper could also serve as a useful first step toward understanding equilibrium behavior in such systems. To the best of our knowledge, the only prior work on the SQ(d) model, $d \geq 2$, for a general class of nonexponential service distributions seems to be the work of Bramson, Lu and Prabhakar [8–11], which focuses on equilibrium behavior rather than dynamical behavior. In particular, under the assumption that the arrival process is Poisson with rate $\lambda < 1$ and the service distribution has a decreasing hazard rate, they show that the stationary distribution of a typical queue in the N-server model converges to a limit, and uncover the interesting phenomenon that when the service distribution is power law, its tail does not always exhibit a doubly exponential decay. However, by the authors' own evaluation (see page 3 of [10]), extending the approach in [10] to more general settings appears to be a difficult problem. The hydrodynamic equations introduced in this article pave the way for an alternative approach to analyzing the equilibrium behavior for a larger class of service distributions and more general, renewal arrivals. Indeed, the hydrodynamic equations have been shown in [1] to have a unique invariant state, which admits a useful computable characterization that provides more detailed information beyond the tail behavior. When combined with the results of [10], this invariant state can be shown to be the limit of the sequence of stationary distributions of the state processes when the service distribution has a decreasing hazard rate function. To establish convergence of the sequence of N-server stationary distributions to this invariant state for the more general class of service distributions considered here, it would suffice to show that the unique invariant state of the hydrodynamic limit is globally attractive. This is a nontrivial interesting problem for future work, but it does provide an alternative approach from that in [10] to the analysis of the equilibrium behavior of large-scale SQ(d) networks.

Finally, we note that the proof of convergence relies on several new techniques, beyond those that have been used in the analysis of scaling limits of measure-valued stochastic process models of queueing networks in other works such as, for example, [22, 24]. First, we introduce a marked point process representation of the dynamics (see Section 5.2) that allows us to prove certain conditional independence properties that are used to identify compensators (with respect to a suitable filtration) of various auxiliary processes that govern the dynamics, such as the cumulative routing and departure processes (see Propositions 5.1 and 5.2). Next, we establish certain renewal estimates to characterize the limit of the scaled compensators. In particular, the compensator of the routing process involves different components whose dynamics evolve on different time scales, and hence the characterization of its limit requires establishing a form of averaging principle (see the Appendix and Sections 6.1.3 and 6.2.2). Furthermore (in Section 6.2), we obtain an alternative dynamical characterization of the solution to the hydrodynamic equations (see Proposition 3.1), which is used to prove relative compactness of the sequence of scaled state processes in Theorem 6.16 and show that any subsequential limit satisfies the hydrodynamic equations. To complete the proof, we establish uniqueness of the solution to the hydrodynamic equations (see Theorem 2.4). The hydrodynamic equations consist of a countable collection of coupled nonlinear measure-valued equations subject to nonstandard boundary conditions that appear to fall outside the class considered in the literature. Two new ingredients that we introduce to facilitate the uniqueness analysis is the nonstandard pseudometric on the space of finite measures defined in (3.1), and a characterization of the evolution of this metric in terms of a certain renewal equation.

1.3. *Outline of the paper.* The rest of the paper is organized as follows. Section 1.4 lists some common notation. Section 2 first introduces the basic assumptions of the model, the state representation and the definition of the hydrodynamic equations, and then states the main results. Section 3 is devoted to the analysis of the hydrodynamic equations, with uniqueness of the solution established in Section 3.1 and an alternative dynamical characterization of the solution obtained in Section 3.2. Section 4 contains a detailed description of the state dynamics in the N-server system. Martingale decompositions for the routing and departure processes are obtained in Section 5.1. The proofs build on a marked point process representation and some conditional independence results established in Section 5.2. Finally, the main convergence results are established in Section 6, with the proof of the convergence result, Theorem 2.6, presented in Section 6.2.3. Proofs of some technical lemmas are relegated to the Appendix.

1.4. *Common notation.* The following notation will be used throughout the paper. We use $\mathbb{Z}$, $\mathbb{Z}_+$ and $\mathbb{N}$ to denote the sets of integers, nonnegative integers and positive integers, respectively. Also, $\mathbb{R}$ is the set of real numbers and $\mathbb{R}_+$ the set of nonnegative real numbers. For $a, b \in \mathbb{R}$, $a \wedge b$ and $a \vee b$ denote the minimum and maximum of a and b, respectively. For a set B, $\mathbb{1}_B(\cdot)$ is the indicator function of the set B (i.e., $\mathbb{1}_B(x) = 1$ if $x \in B$ and $\mathbb{1}_B(x) = 0$ otherwise). When B is a measurable subset of a probability space $(\Omega, \mathbb{F})$, we omit the explicit dependence on ω and write $\mathbb{1}_B(\omega)$ as $\mathbb{1}_B$. Moreover, with a slight abuse of notation, on every domain V, $\mathbf{1}$ denotes the constant function equal to 1 on V. Also, $\mathbf{Id}$ is the identity function on $[0,\infty)$, that is, $\mathbf{Id}(t)=t$, for all $t\geq 0$.

For a topological space V, we let $\mathbb{C}(V)$, $\mathbb{C}_b(V)$ and $\mathbb{C}_c(V)$ be, respectively, the space of continuous functions, bounded continuous functions, and continuous functions with compact support on V. For $f\in\mathbb{C}_b(V)$, $\|f\|_\infty$ denotes $\sup_{s\in V}|f(s)|$. When $V=[0,\infty)$, for $T\geq 0$, $\|f\|_T$ denotes $\sup_{s\in[0,T]}|f(s)|$, and recall that $w_f(\delta,T):=\sup\{|f(t)-f(s)|; s,t\in[0,T], |s-t|\leq\delta\}$, $\delta>0$, is the modulus of continuity of f on the interval $[0,T]$. For $V=[0,L)$, $L\in[0,\infty]$, $\mathbb{C}_b^1(V)$ is the set of functions $f\in\mathbb{C}_b(V)$ for which the first derivative, denoted by f', exists and is bounded and continuous on V. Similarly, when $V\subset\mathbb{R}^2$ is the product of two intervals in $\mathbb{R}$, $\mathbb{C}_b^{1,1}(V)$ (resp., $\mathbb{C}_c^{1,1}(V)$) is the set of functions $(x,s)\mapsto\varphi(x,s)$ in $\mathbb{C}_b(V)$ (resp., $\mathbb{C}_c(V)$) for which the first order partial derivatives φ_x and φ_s exist and are bounded and continuous (resp., continuous with compact support) on V. Also, let $\mathbb{AC}(V)$ denote the space of real-valued functions that are absolutely continuous on every bounded subset of V.

For a metric space $\mathbb{X}$, $\mathbb{D}_{\mathbb{X}}[0,\infty)$ is the set of $\mathbb{X}$-valued functions on $[0,\infty)$ that are right continuous and have finite left limits on $(0,\infty)$, and $\mathbb{C}_{\mathbb{X}}[0,\infty)$ is the subset of continuous functions on $[0,\infty)$. For every function $f\in\mathbb{D}_{\mathbb{X}}[0,\infty)$ and $T\geq 0$, $w'(f,\delta,T)$ is the modulus of continuity of f in $\mathbb{D}_{\mathbb{X}}[0,\infty)$; see equation (3.6.2) of [17] for a precise definition of w'. Furthermore, for every function $f\in\mathbb{D}_{\mathbb{R}}[0,\infty)$, we define

$$[f]_t := \lim_{|\pi|\to 0}\sum_{k=1}^{n}\big(f(t_k)-f(t_{k-1})\big)^2,$$

where the limit is taken over all partitions $\pi=\{t_0=0,t_1,\dots,t_n=t\}$ $[0,t]$ with $|\pi|:=\max_{k=1,\dots,n}|t_k-t_{k-1}|$. When f is a càdlàg stochastic process, the limit is defined in the sense of convergence in probability.

Finally, $\mathbb{L}^1(0,\infty)$, $\mathbb{L}^2(0,\infty)$ and $\mathbb{L}^\infty(0,\infty)$, denote, respectively, the spaces of integrable, square-integrable and essentially bounded functions on $(0,\infty)$, equipped with their corresponding standard norms. Also, $\mathbb{L}^1_{\text{loc}}(0,\infty)$ denotes the space of locally integrable functions on $[0,\infty)$. For any $f\in\mathbb{L}^1_{\text{loc}}(0,\infty)$ and a function g that is bounded on finite intervals, $g*f$ denotes the (one-sided) convolution of the two functions, defined as $f*g(t):=\int_0^t f(t-s)g(s)\,ds$, $t\geq 0$.

For every subset V of $\mathbb{R}$ or $\mathbb{R}^2$ endowed with the Borel sigma-algebra, let $\mathbb{M}_F(V)$ (resp., $\mathbb{M}_{\leq 1}(V)$) be the space of finite positive (resp., sup-probability) measures on V. For $\mu\in\mathbb{M}_F(V)$ and any bounded Borel-measurable function f on V, we denote the integral of f with respect to μ by

$$\langle f,\mu\rangle := \int_V f(x)\mu(dx).$$

Given $\mu\in\mathbb{M}_F(V)$ and a function f defined on a larger set $\tilde{V}\supseteq V$, by some abuse of notation, we will write $\langle f,\mu\rangle$ to denote $\langle f_{|V},\mu\rangle=\int_V f(x)\mu(dx)$, where

$f_{|V}$ denotes the restriction of f to V. For every measure μ with representation $\mu = \mu^+ - \mu^-; \mu^+, \mu^- \in \mathbb{M}_f[0,\infty)$, we extend the bracket notation by setting $\langle f, \mu\rangle := \langle f, \mu^+\rangle - \langle f, \mu^-\rangle$. We equip $\mathbb{M}_F(V)$ and $\mathbb{M}_{\leq 1}(V)$ with the weak topology: $\mu_n \Rightarrow \mu$ if and only if $\langle f, \mu_n\rangle \to \langle f, \mu\rangle$ for all $f \in \mathbb{C}_b[0,\infty)$. Recall that the Prohorov metric d_P on $\mathbb{M}_F(V)$ (page 72 of [7]) induces the same topology (see Theorem 6.8 of Chapter 1 in [7]).

Also, we denote by $\mathbb{M}(V)$ the space of Radon measures on V, that is, the space of measures on V that assign finite mass to every relatively compact subset of V. Alternatively, one can identify a Radon measure $\mu \in \mathbb{M}(V)$ with a linear functionals $\varphi \mapsto \mu(\varphi) := \int_V \varphi(x)\mu(dx)$ on the space $\mathbb{C}_c(V)$ of compactly supported functions on V such that for every compact set $\mathcal{K} \subset V$, there exists a finite $C_{\mathcal{K}}$ such that

$$\mu(\varphi) \leq C_{\mathcal{K}}\|\varphi\|_\infty, \qquad \forall\varphi \in \mathbb{C}_c(V) \text{ with } \operatorname{supp}(\varphi) \subset \mathcal{K}.$$

2. Main results.

2.1. *Basic assumptions.* Consider the SQ(d) model with N servers described in the Introduction. For $t \geq 0$, let $E^{(N)}(t)$ denote the number of jobs that arrived to the network in the interval $[0,t]$. We start by stating our assumptions on the cumulative arrival process $E^{(N)}$. Let $\widetilde{E}$ be a delayed renewal process with interarrival times $\widetilde{u}_n, n \geq 1$, whose cumulative distribution function $G_{\widetilde{E}}$ has a density $g_{\widetilde{E}}$ and mean λ^{-1}, for some $\lambda > 0$, and delay $\widetilde{u_0}$ that satisfies $\mathbb{P}\{\widetilde{u}_0 > r\} = \overline{G}_{\widetilde{E}}(\tilde{R}+r)/\overline{G}_{\widetilde{E}}(\tilde{R})$, for some $\tilde{R} \geq 0$, where $\overline{G}_{\widetilde{E}} := 1 - G_{\widetilde{E}}$.

ASSUMPTION I. The arrival process satisfies $E^{(N)}(t) = \widetilde{E}(Nt), t \geq 0$.

Note that Assumption I implies that $E^{(N)}$ is a delayed renewal process with delay $u_0^{(N)} := \widetilde{u}_0/N$, and interarrival times $u_n^{(N)} := \widetilde{u}_n/N$, $n \geq 1$, with common distribution $G_E^{(N)}(x) := G_{\widetilde{E}}(Nx)$, $x \geq 0$, and probability density function $g_E^{(N)}(\cdot) = Ng_{\widetilde{E}}(N\cdot)$. Moreover, setting $R^{(N)} := \widetilde{R}/N$, we have

$$\mathbb{P}\{u_0^{(N)} > r\} = \frac{\overline{G}_E^{(N)}(R^{(N)} + r)}{\overline{G}_E^{(N)}(R^{(N)})}, \qquad r \geq 0. \tag{2.1}$$

For future purposes, we also define the backward recurrence time of $E^{(N)}$:

$$R_E^{(N)}(t) := \begin{cases} R^{(N)} + t & \text{if } 0 \leq t < u_0, \\ t - \sup\{s \geq 0, E^{(N)}(s) < E^{(N)}(t)\} & \text{if } t \geq u_0, \end{cases} \tag{2.2}$$

where in this particular definition, the supremum of an empty set should be interpreted as zero. Note that $R_E^{(N)}(0) = R^{(N)}$.

Next, let G denote the cumulative distribution function of the i.i.d. service times $\{v_j; j \in \mathbb{Z}\}$, and let $\overline{G} := 1 - G$. We impose the following conditions on G.

ASSUMPTION II. The service time distribution G has the following properties:

(a) G has a density g and finite mean which can (and will) be set to 1.
(b) There exists $\ell_0 < L$ such that the hazard rate function

$$h(x) := \frac{g(x)}{\overline{G}(x)}, \qquad x \in [0, L), \tag{2.3}$$

where $L := \sup\{x \in [0,\infty) : G(x) < 1\}$, is either bounded or lower semicontinuous on (ℓ_0, L).
(c) The density g is bounded on every finite interval of $[0,\infty)$.

Note that both Assumptions II(b) and II(c) hold if either g is continuous or h is bounded on $[0,\infty)$.

2.2. *State representation.* Recall from the Introduction that $\nu_\ell^{(N)}(t)$ is a (random) finite measure on $[0,\infty)$ that has a unit delta mass at the age (i.e., amount of time spent in service) of each job that, at time t, is in service at a queue of length no less than ℓ. Since the maximum number of jobs in service at any time is N, $\nu_\ell^{(N)}(t)/N$ takes values in the space $\mathbb{M}_{\leq 1}[0,L)$ of subprobability measures on $[0,L)$. The state of the system at time t will be represented by $\nu^{(N)}(t) := (\nu_\ell^{(N)}(t); \ell \geq 1)$. The scaled state $\nu^{(N)}(t)/N$ takes values in the space

$$\begin{aligned}\mathbb{S} := \big\{&(\mu_\ell; \ell \geq 1) \in \mathbb{M}_{\leq 1}[0,L)^{\mathbb{N}}; \langle f, \mu_\ell\rangle \geq \langle f, \mu_{\ell+1}\rangle,\\ &\forall \ell \geq 1, f \in \mathbb{C}_b[0,\infty), f \geq 0\big\}\end{aligned} \tag{2.4}$$

of ordered sequences of subprobability measures. We equip $\mathbb{S}$ with the metric

$$d_{\mathbb{S}}(\mu, \tilde{\mu}) := \sup_{\ell \geq 1} \frac{d_P(\mu_\ell, \tilde{\mu}_\ell)}{\ell}, \tag{2.5}$$

where d_P is the Prohorov metric. Thus, a sequence $\{\mu^n\}$ converges to μ in $\mathbb{S}$ if and only if for every $\ell \geq 1$, $\{\mu_\ell^n\}$ converges weakly to μ_ℓ. Recall that weak convergence and the Prohorov metric are defined in Section 1.4.

Recall that $\mathbf{1}$ denotes the function that is identically one, and note that

$$S_\ell^{(N)}(t) := \langle \mathbf{1}, \nu_\ell^{(N)}(t)\rangle, \qquad t \geq 0, \ell \geq 1, \tag{2.6}$$

is the number of queues with length at least ℓ at time t. Moreover, let $X^{(N)}(t)$ be the total number of jobs in the system at time t (including those in service and those waiting in queue). Since $S_\ell^{(N)}(t) - S_{\ell+1}^{(N)}(t)$ is the number of queues with length exactly ℓ, we have

$$X^{(N)}(t) = \sum_{\ell \geq 1} [\ell(S_\ell^{(N)}(t) - S_{\ell+1}^{(N)}(t))] = \sum_{\ell \geq 1} S_\ell^{(N)}(t) = \sum_{\ell \geq 1} \langle \mathbf{1}, \nu_\ell^{(N)}(t)\rangle. \tag{2.7}$$

Note that the right-hand side above is in fact a finite sum if and only if the number of jobs in the system is finite at time t, which in turn holds if and only if that condition holds for $t = 0$ since $X^{(N)}(t) \le X^{(N)}(0) + E^{(N)}(t)$. Finally, for $t \ge 0$ and $\ell \in \mathbb{N}$, let $D_\ell^{(N)}(t)$ denote the total number of jobs that completed service in the interval $[0, t]$ at a queue that had length ℓ just prior to service completion. All these random elements are assumed to be supported on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

2.3. *Hydrodynamic equations.* We now introduce the hydrodynamic equations, which will be shown to characterize the "functional law of large numbers" or "fluid" limit of the state of the network. The terminology refers to the fact that we are looking at the limiting dynamics of the empirical measure of an interacting particle system. For $d \ge 2$, let

$$\mathfrak{P}_d(x, y) := \frac{x^d - y^d}{x - y} = \sum_{m=0}^{d-1} x^m y^{d-1-m}. \tag{2.8}$$

When $d = 2$, we have the simple form $\mathfrak{P}_2(x, y) = x + y$, and in general, for $x, y, \tilde{x}, \tilde{y} \le 1$,

$$\mathfrak{P}_d(x, y) \le d \quad \text{and} \quad \mathfrak{P}_d(x, y) - \mathfrak{P}_d(\tilde{x}, \tilde{y}) \le d^2\big((x - \tilde{x}) + (y - \tilde{y})\big). \tag{2.9}$$

DEFINITION 2.1 (Hydrodynamic equations). Given $\lambda > 0$ and $\nu(0) \in \mathbb{S}$, $\{\nu(t) = (\nu_\ell(t); \ell \ge 1); t \ge 0\}$ in $\mathbb{C}_{\mathbb{S}}[0, \infty)$ is said to solve the *hydrodynamic equations* associated with $(\lambda, \nu(0))$ if and only if for every $t \in [0, \infty)$,

$$\int_0^t \langle h, \nu_1(s)\rangle\, ds < \infty, \tag{2.10}$$

and for every $\ell \ge 1$,

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle\, ds - D_\ell(t), \tag{2.11}$$

where

$$D_\ell(t) := \int_0^t \langle h, \nu_\ell(s)\rangle\, ds, \qquad \forall \ell \ge 1, \tag{2.12}$$

and for every $f \in \mathbb{C}_b[0, \infty)$,

$$\begin{aligned} \langle f, \nu_\ell(t)\rangle &= \left\langle f(\cdot + t)\frac{\overline{G}(\cdot + t)}{\overline{G}(\cdot)}, \nu_\ell(0)\right\rangle + \int_{[0,t]} f(t - s)\overline{G}(t - s)\, dD_{\ell+1}(s) \\ &\quad + \int_0^t \left\langle f(\cdot + t - s)\frac{\overline{G}(\cdot + t - s)}{\overline{G}(\cdot)}, \eta_\ell(s)\right\rangle ds, \end{aligned} \tag{2.13}$$

with

$$\eta_\ell(t) := \begin{cases} \lambda(1 - \langle \mathbf{1}, \nu_1(t)\rangle^d)\delta_0 & \text{if } \ell = 1, \\ \lambda\mathfrak{P}_d(\langle \mathbf{1}, \nu_{\ell-1}(t)\rangle, \langle \mathbf{1}, \nu_\ell(t)\rangle)(\nu_{\ell-1}(t) - \nu_\ell(t)) & \text{if } \ell \ge 2. \end{cases} \tag{2.14}$$

Given a solution ν to the hydrodynamic equations, we define

$$S_\ell(t) := \langle \mathbf{1}, \nu_\ell(t) \rangle, \qquad t \geq 0, \ell \geq 1. \tag{2.15}$$

REMARK 2.2. The bound (2.10) implies that for every $\ell \geq 1$, the process D_ℓ is well defined.

REMARK 2.3. Definition 2.1 of the hydrodynamic equations and the corresponding uniqueness result in Theorem 2.4 can be generalized to time-varying rates by simply replacing the constant arrival λ everywhere with a nonnegative locally integrable function $\lambda(\cdot)$.

We now state our first main result, which is proved in Section 3.1.

THEOREM 2.4. *Suppose Assumptions* I *and* II(a) *hold. Then, for every* $\lambda > 0$ *and* $\nu(0) \in \mathbb{S}$, *the hydrodynamic equations associated with* $(\lambda, \nu(0))$ *have at most one solution.*

We now provide some intuition into the form of the hydrodynamic equations. Given $a(s)$, the age of a job in service at time s, the mean conditional probability that this job will complete service in the time interval $(s, s + ds)$ is roughly $h(a(s))\,ds$. Summing over the ages of all jobs in service at queues of length no less than ℓ, we see that the conditional mean departure rate from such queues at time s is $\langle h, \nu_\ell(s)\rangle$. In the large N limit, the scaled departure process coincides with its mean, thus giving rise to the equality in (2.12). Next, to understand the mass balance equation (2.11), fix $\ell \geq 1$ and note that in analogy with (2.6), $\langle \mathbf{1}, \nu_\ell(t)\rangle$ represents the limit fraction of queues of length no less than ℓ at time t. Over the interval $[0, t]$, this quantity decreases due to departures from queues of length precisely ℓ, which is given by $D_\ell(t) - D_{\ell+1}(t)$, and increases due to exogenous arrivals to queues of length $\ell - 1$. To quantify the latter, note that λ is the scaled mean arrival rate of jobs to the network and the probability that an arriving job is routed to a queue of length $\ell - 1$ at time s is approximately equal to $(\langle \mathbf{1}, \nu_{\ell-1}(s)\rangle)^d - (\langle \mathbf{1}, \nu_\ell(s)\rangle)^d$, which is equal to $\mathfrak{P}_d(\langle \mathbf{1}, \nu_{\ell-1}(s)\rangle, \langle \mathbf{1}, \nu_\ell(s)\rangle)(\langle \mathbf{1}, \nu_{\ell-1}(s)\rangle - \langle \mathbf{1}, \nu_\ell(s)\rangle)$, with the convention $\langle \mathbf{1}, \nu_0(s)\rangle := 1$. Thus, with η_ℓ defined by (2.14), $\langle \mathbf{1}, \eta_\ell(s)\rangle$ represents the scaled arrival rate at time s of jobs to queues of length $\ell - 1$, and $\int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle\, ds$ represents the total exogenous arrivals to such queues over the interval $[0, t]$. These observations, when combined, justify the form of (2.11).

Equation (2.13) is a more involved mass balance equation, whose right-hand side consists of three terms that contribute to the measure $\nu_\ell(t)$. The first term on the right-hand side accounts for jobs already in service at time 0. Any such job, conditioned on having initial age $a(0)$, would still be in service at time t, with age $a(0) + t$, with probability $\overline{G}(a(0) + t)/\overline{G}(a(0))$. The second term represents the contribution to $\nu_\ell(t)$ due to jobs that entered service at some time $s \in (0, t]$ at a queue of length no less than ℓ at time s and that are still in service at time t. Such service entries occur due to departures of jobs at time s from a queue no less than $\ell + 1$ prior to departure, which would happen at rate $dD_{\ell+1}(s)$. Further, the job entering service would have age 0 at time s and so would still be in service at time t (with age $t - s$) with probability $\overline{G}(t - s)$. Finally, the last term captures the contribution due to jobs that were in service at a queue of length $\ell - 1$ at some time $s \in [0, t]$ when its length increased by one due to the routing of a job to that queue. If $\ell > 1$, and $a(s)$ was the age of the job in service at that queue at time s, the job would still be in service at time t only if its service time were greater than $a(s) + t - s$ (given that it was clearly greater than $a(s)$), which has probability $\overline{G}(a(s) + t - s)/\overline{G}(a(s))$. Now, the (limit) distribution of ages in service at queues of length $\ell - 1$ at time s is $\nu_{\ell-1}(s) - \nu_\ell(s)$. When multiplied by the (limit) rate at which jobs are routed to a random queue of length $\ell - 1$, which is $\lambda\mathfrak{P}_d(\langle\mathbf{1}, \nu_{\ell-1}(s)\rangle, \langle\mathbf{1}, \nu_\ell(s)\rangle)$, yields $\eta_\ell(s)$. The case $\ell = 1$ can be argued similarly. This explains the form of the third term on the right-hand side of (2.13). The above discussion also suggests why the limit of a more general class of routing algorithms could be characterized similarly, but with a suitably modified definition of η_ℓ.

2.4. *Convergence result.* For $H = E, D, \nu_\ell, \nu, S_\ell$, we define the scaled version of $H^{(N)}$ as follows:

$$\overline{H}^{(N)}(t) = \frac{H^{(N)}(t)}{N}, \qquad N \in \mathbb{N}, t \geq 0. \tag{2.16}$$

The following condition is imposed on the initial state of the network.

ASSUMPTION III. The sequence of initial conditions satisfies the following:

(a) For every $N \in \mathbb{N}$, $X^{(N)}(0) = \sum_{\ell\geq 1}\langle\mathbf{1}, \nu_\ell^{(N)}(0)\rangle < \infty$ almost surely, $E^{(N)}$ and the random queue choices in the load balancing algorithm are independent of $\nu^{(N)}(0)$, and for each job j that is in service at time 0, its service time v_j is conditionally independent of $\nu^{(N)}(0)$ given its initial age $a_j(0)$; see (4.4) for further details.

(b) There exists $\nu(0) = (\nu_\ell(0); \ell \geq 1) \in \mathbb{S}$ such that $\overline{\nu}^{(N)}(0) \to \nu(0)$ in $\mathbb{S}$, $\mathbb{P}$-almost surely, as $N \to \infty$.

(c) $\limsup_N \mathbb{E}[\overline{X}^{(N)}(0)] < \infty$, and $\overline{X}^{(N)}(0) \to X(0)$ as $N \to \infty$, where $X(0) := \sum_{\ell\geq 1}\langle\mathbf{1}, \nu_\ell(0)\rangle$.

We now state some immediate consequences of Assumptions I and III(c).

LEMMA 2.5. *Suppose Assumption* I *holds. Then, as* $N \to \infty$, $\overline{E}^{(N)} \to \lambda\mathbf{Id}$ *in* $\mathbb{D}_\mathbb{R}[0, \infty)$, $\mathbb{P}$-*almost surely. Moreover, for all* $t \geq 0$, $\mathbb{E}[\overline{E}^{(N)}(t)] \to \lambda t$ *as* $N \to \infty$, *and hence*,

$$\limsup_{N\to\infty}\mathbb{E}\bigl[\overline{E}^{(N)}(t)\bigr]<\infty. \tag{2.17}$$

In addition, *if Assumption* III(c) *holds*, *then for every* $t\geq 0$,

$$\limsup_{N\to\infty}\mathbb{E}\bigl[\overline{X}^{(N)}(0)+\overline{E}^{(N)}(t)\bigr]<\infty. \tag{2.18}$$

PROOF. The almost-sure convergence of $\overline{E}^{(N)}$ to $\lambda\mathbf{Id}$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ follows from Assumption I and the functional law of large numbers for renewal processes (e.g., see Theorem 5.10 of [14]). Also, for $t\geq 0$, by the elementary renewal theorem (e.g., see Proposition V.1.4 of [6]), $\lim_{N\to\infty}\mathbb{E}[\overline{E}^{(N)}(t)] = \lim_{N\to\infty}\widetilde{E}(Nt)/N=\lambda t$, where $\widetilde{E}$ is the delayed renewal process of Assumption I. This implies (2.17), which along with Assumption III(c), implies (2.18). □

We now state the second main result, whose proof is given in Section 6.2.3.

THEOREM 2.6. *Suppose Assumptions* I–III *hold. Then there exists a unique solution* $\nu\in\mathbb{C}_{\mathbb{S}}$ *to the hydrodynamic equations associated with* $(\lambda,\nu(0))$, *and the sequence* $\{\overline{\nu}^{(N)}\}$ *converges in distribution to* ν.

REMARK 2.7. Theorem 2.6 implies that for every $t\geq 0$, the (fluid-scaled) number of jobs in the system $\overline{X}^{(N)}(t)$ converges in distribution to $X(t)=\sum_{\ell\geq 1}\langle\mathbf{1},\nu_\ell(t)\rangle$. This follows from the dominated convergence theorem and the fact that when $X(0)=\sum_{\ell\geq 1}\langle\mathbf{1},\nu_\ell(0)\rangle$ is finite, $X(t)$ remains finite for all $t\geq 0$, that is, the solutions to the hydrodynamic equations preserves the set

$$\Bigl\{(\mu_\ell;\ell\geq 1)\in\mathbb{S};\sum_{\ell\geq 1}\langle\mathbf{1},\mu_\ell\rangle<\infty\Bigr\}.$$

To see this fact, note that for every $L\in\mathbb{N}$, by summing over (2.11) and by the definition (2.14) of η, we have

$$\begin{aligned}\sum_{\ell=1}^{L}\langle\mathbf{1},\nu_\ell(t)\rangle&=\sum_{\ell=1}^{L}\langle\mathbf{1},\nu_\ell(0)\rangle+D_{L+1}(t)-D_1(t)+\lambda\int_0^t\bigl(1-\langle\mathbf{1},\nu_1(t)\rangle^d\bigr)\,ds\\&\leq X(0)+\lambda t.\end{aligned}$$

Letting $L\to\infty$ implies $X(t)\leq X(0)+\lambda t$.

As a corollary, we establish a "propagation of chaos" result, whose proof is also deferred to Section 6.2.3. Let $X^{(N),i}(t)$ be the length of the ith queue at time t, and if the queue is initially nonempty, let $a^{(N),i}(0)$ be the initial age of the job receiving service at the ith queue.

COROLLARY 2.8. *Suppose Assumptions I–III hold, and the initial conditions are exchangeable, that is, for every N and any permutation π of the queue indices $\{1,\dots,N\}$, the random vector*

$$\big(X^{(N),\pi(i)}(0), a^{(N),\pi(i)}(0)\mathbb{1}_{\{X^{(N),\pi(i)}(0)>0\}}; i=1,\dots,N\big),$$

has the same distribution. Let ν be the solution to the hydrodynamic equations associated with $(\lambda,\nu(0))$ and let $\{S_\ell,\ell\ge 1\}$ be as defined in (2.15). Then, for every $\ell\ge 1$ and $t\ge 0$,

$$\lim_{N\to\infty}\mathbb{P}\{X^{(N),1}(t)\ge \ell\} = S_\ell(t), \tag{2.19}$$

and for ever fixed $k\ge 0$ and $\ell_1,\dots,\ell_k\in\mathbb{N}$,

$$\lim_{N\to\infty}\mathbb{P}\{X^{(N),1}(t)\ge \ell_1,\dots,X^{(N),k}(t)\ge\ell_k\} = \prod_{m=1}^{k} S_{\ell_m}(t). \tag{2.20}$$

3. Analysis of the hydrodynamic equations.

In Section 3.1, we prove Theorem 2.4. In Section 3.2, we obtain a dynamical characterization of the hydrodynamic equations in terms of a measure-valued PDE, which is used in Section 6.2.3 to prove Theorem 2.6.

3.1. *Proof of uniqueness of the solution to the hydrodynamic equations.*

PROOF OF THEOREM 2.4. Fix $\nu(0)\in\mathbb{S}$, $\lambda>0$, and let ν and $\tilde\nu$ both be solutions to the hydrodynamic equations associated with $(\lambda,\nu(0))$, and let $\tilde D,\tilde\eta,\tilde S$ be defined as in (2.12)–(2.15), but with ν replaced by $\tilde\nu$. For $\ell\ge 1$, define $\Delta H_\ell := H_\ell - \tilde H_\ell$ for $H=\nu, D,\eta, S$. Consider the parameterized family of continuous bounded functions

$$\mathbb{F} := \left\{\vartheta^r := \frac{\overline{G}(\cdot+r)}{\overline{G}(\cdot)}; r\ge 0\right\}\subset \mathbb{C}_b[0,L)\cap\mathbb{AC}[0,\infty),$$

where the last inclusion holds by Assumption II. Note that $\mathbf{1}=\vartheta^0\in\mathbb{F}$, and for every $\ell\ge 1$ and $t\ge 0$, define

$$V_\ell(t) := \sup_{f\in\mathbb{F}}\big|\langle f,\Delta\nu_\ell(t)\rangle\big|. \tag{3.1}$$

By (2.14) with $\ell=1$, (2.15) and (2.8), for $s\ge 0$ and every $f\in\mathbb{C}_b[0,L)$ we have

$$\begin{aligned}\langle f,\Delta\eta_1(s)\rangle &= \lambda f(0)\big((\tilde S_1(s))^d - (S_1(s))^d\big)\\ &= -\lambda f(0)\mathfrak{P}_d\big(\tilde S_1(s),S_1(s)\big)\langle \mathbf{1},\Delta\nu_1(s)\rangle.\end{aligned}\tag{3.2}$$

Therefore, for every $f\in\mathbb{F}$, since $\mathfrak{P}_d(x,y)\le d$ and $\|f\|_\infty\le 1$,

$$\big|\langle f,\Delta\eta_1(s)\rangle\big| \le \lambda d\big|\langle\mathbf{1},\Delta\nu_1(s)\rangle\big| \le \lambda d V_1(s),\qquad f\in\mathbb{F}. \tag{3.3}$$

Next, for $\ell \geq 2$, $s \geq 0$ and $f \in \mathbb{C}_b[0, L)$, again invoking (2.14), (2.15) and (2.8), we have

$$\begin{aligned} \langle f, \Delta\eta_\ell(s)\rangle = {} & \lambda\mathfrak{P}_d(S_{\ell-1}(s), S_\ell(s))\langle f, \Delta\nu_{\ell-1}(s) - \Delta\nu_\ell(s)\rangle \\ & + \lambda(\mathfrak{P}_d(S_{\ell-1}(s), S_\ell(s)) - \mathfrak{P}_d(\tilde{S}_{\ell-1}(s), \tilde{S}_\ell(s))) \\ & \times \langle f, \tilde{\nu}_{\ell-1}(s) - \tilde{\nu}_\ell(s)\rangle. \end{aligned} \tag{3.4}$$

Hence, for $f \in \mathbb{F}$, given $\|f\|_\infty \leq 1$, $\nu_{\ell-1} \geq \nu_\ell$, $\tilde{\nu}_{\ell-1} \geq \tilde{\nu}_\ell$, $\nu_\ell, \tilde{\nu}_\ell \in \mathcal{M}_{\leq 1}[0, L)$, $\mathfrak{P}_d(x, y) \leq d$ and inequality (2.9), we have

$$|\langle f, \Delta\eta_\ell(s)\rangle| \leq \lambda(d + d^2)(V_{\ell-1}(s) + V_\ell(s)), \qquad f \in \mathbb{F}. \tag{3.5}$$

Now, for $f \in \mathbb{C}_b[0, \infty) \cap \mathbb{AC}[0, \infty)$ applying integration by parts to (2.13), we obtain

$$\begin{aligned} \langle f, \nu_\ell(t)\rangle = {} & \left\langle f(\cdot + t)\frac{\overline{G}(\cdot + t)}{\overline{G}(\cdot)}, \nu_\ell(0)\right\rangle + f(0)D_{\ell+1}(t) \\ & + \int_{[0,t]} (f\overline{G})'(t - s)D_{\ell+1}(s)\,ds \\ & + \int_0^t \left\langle f(\cdot + t - s)\frac{\overline{G}(\cdot + t - s)}{\overline{G}(\cdot)}, \eta_\ell(s)\right\rangle ds. \end{aligned} \tag{3.6}$$

Also, for every $t \geq s \geq 0$, $r \geq 0$, we have $\vartheta^r(0) = \overline{G}(r)$, $(\vartheta^r\overline{G})'(t - s) = -g(t - s + r)$, and

$$\vartheta^r(x + t - s)\frac{\overline{G}(x + t - s)}{\overline{G}(x)} = \frac{\overline{G}(x + t - s + r)}{\overline{G}(x)} = \vartheta^{t-s+r}(x), \qquad x \in [0, L).$$

Thus, substituting $f = \vartheta^r$ in (3.6), both as is and when ν_ℓ, D_ℓ, η_ℓ are replaced by $\tilde{\nu}_\ell$, $\tilde{D}_\ell$, $\tilde{\eta}_\ell$, respectively, and recalling $\Delta\nu_\ell(0) = 0$, we have

$$\begin{aligned} \langle\vartheta^r, \Delta\nu_\ell(t)\rangle = {} & \overline{G}(r)\Delta D_{\ell+1}(t) - \int_0^t g(t - s + r)\Delta D_{\ell+1}(s)\,ds \\ & + \int_0^t \langle\vartheta^{t-s+r}, \Delta\eta_\ell(s)\rangle\,ds. \end{aligned} \tag{3.7}$$

Since $\vartheta^0 = \mathbf{1}$, equation (3.7) for $r = 0$ gives

$$\begin{aligned} \langle\mathbf{1}, \Delta\nu_\ell(t)\rangle = {} & \Delta D_{\ell+1}(t) - \int_0^t g(t - s)\Delta D_{\ell+1}(s)\,ds \\ & + \int_0^t \langle\vartheta^{t-s}, \Delta\eta_\ell(s)\rangle\,ds. \end{aligned} \tag{3.8}$$

Since $\{\nu_\ell\}_{\ell\in\mathbb{N}}$ and $\{\tilde{\nu}_\ell\}_{\ell\in\mathbb{N}}$ satisfy (2.11) and its analog, and that $\Delta\nu_\ell(0) = 0$, we have

$$\Delta D_{\ell+1}(t) = \langle\mathbf{1}, \Delta\nu_\ell(t)\rangle + \Delta D_\ell(t) - \int_0^t \langle\mathbf{1}, \Delta\eta_\ell(s)\rangle\,ds. \tag{3.9}$$

Combining (3.9) and (3.8), it follows that for $\ell \geq 2$, ΔD_ℓ satisfies the renewal equation

$$\Delta D_\ell(t) = g * \Delta D_\ell(t) + F_\ell(t),$$

with

$$F_\ell(t) := \int_0^t \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle\, ds + g * \langle \mathbf{1}, \Delta\nu_\ell\rangle(t) - \left(g * \int_0^{\cdot} \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle\, ds\right)(t)$$
$$- \int_0^t \langle \vartheta^{t-s}, \Delta\eta_\ell(s)\rangle\, ds.$$

Since $\Delta\nu_\ell$ is the difference of two measures in $\mathbb{D}_{\mathbb{M}_F[0,L)}[0,\infty)$, $\langle \mathbf{1}, \Delta\nu_\ell(\cdot)\rangle$ and $\langle f, \Delta\eta_\ell(\cdot)\rangle$, $f \in \mathbb{C}_b[0,\infty)$, are also locally integrable, and hence F_ℓ is uniformly bounded on finite intervals (i.e., $\|F\|_t < \infty$ for all $t \geq 0$). Moreover, (2.10) ensures that ΔD_ℓ is also bounded on finite intervals. Therefore, by Theorem V.2.4 of [6],

$$\Delta D_\ell(t) = F_\ell(t) + u * F_\ell(t), \tag{3.10}$$

where u is the renewal density of G on $[0, L)$. Note that since G has density g, by Proposition V.2.7 of [6], u exists and satisfies the equation $u = u*g+g$. Moreover, since g is locally bounded due to Assumption II(c), u is bounded on every finite interval of $[0, L)$ by another application of Theorem V.2.4 of [6]. Substituting the definition of F_ℓ into equation (3.10) and using the relation $u * g + g = u$, we have

$$\Delta D_\ell(t) = \int_0^t \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle\, ds + u * \langle \mathbf{1}, \Delta\nu_\ell\rangle(t) - \int_0^t \langle \vartheta^{t-s}, \Delta\eta_\ell(s)\rangle\, ds$$
$$- \left(u * \int_0^{\cdot} \langle \vartheta^{\cdot-s}, \Delta\eta_\ell(s)\rangle\, ds\right)(t). \tag{3.11}$$

Next, we bound each term on the right-hand side of (3.11). Fix $T \geq 0$. Then (3.5) implies

$$\left|\int_0^t \langle \mathbf{1}, \Delta\eta_\ell(s)\rangle\, ds\right| \leq \lambda(d+d^2)\int_0^t \big(V_{\ell-1}(s) + V_\ell(s)\big)\, ds, \qquad t \leq T. \tag{3.12}$$

Moreover, recalling the notation $\|f\|_T = \sup_{t\in[0,T]} |f(t)|$, we also have

$$\big|u * \langle \mathbf{1}, \Delta\nu_\ell\rangle(t)\big| \leq \int_0^t u(t-s)\big|\langle \mathbf{1}, \Delta\nu_\ell(s)\rangle\big|\, ds$$
$$\leq \|u\|_T \int_0^t V_\ell(s)\, ds, \qquad t \leq T. \tag{3.13}$$

Furthermore, for the function $\zeta_\ell(t) := \int_0^t \langle \vartheta^{t-s}, \Delta\eta_\ell(s)\rangle\, ds$, the bound (3.5) with $f = \vartheta^{t-s}$ implies

$$|\zeta_\ell(s)| \leq \lambda(d+d^2)\int_0^s \big(V_{\ell-1}(v) + V_\ell(v)\big)\, dv, \qquad s \geq 0, \tag{3.14}$$

and hence, applying Tonelli's theorem in the third inequality below, we obtain

$$
\begin{aligned}
|u * \zeta_\ell(t)| &\leq \int_0^t u(t-s)|\zeta_\ell(s)|\,ds \\
&\leq \lambda(d+d^2)\int_0^t\int_0^s u(t-s)\big(V_{\ell-1}(v)+V_\ell(v)\big)\,dv\,ds \\
&\leq \lambda(d+d^2)\int_0^t \big(V_{\ell-1}(v)+V_\ell(v)\big)\Big(\int_v^t u(t-s)\,ds\Big)\,dv \\
&\leq \lambda(d+d^2)U(T)\int_0^t \big(V_{\ell-1}(s)+V_\ell(s)\big)\,ds, \qquad \forall t \leq T, \qquad (3.15)
\end{aligned}
$$

where $U(\cdot) = 1 + \int_0^{\cdot} u(s)\,ds$ is the renewal function associated with G. From equation (3.11) and the bounds (3.12)–(3.15), for $\ell \geq 2$ we then have

$$
\|\Delta D_\ell\|_t \leq C(T)\int_0^t \big(V_{\ell-1}(s)+V_\ell(s)\big)\,ds, \qquad \forall t \leq T, \tag{3.16}
$$

with $C(T) := \|u\|_T + \lambda(d+d^2)(2+U(T)) < \infty$. Finally, incorporating (3.3), (3.5) and (3.16), with ℓ replaced by $\ell+1$, into (3.7), we have

$$
|\langle f, \Delta\nu_1(t)\rangle| \leq 3C(T)\int_0^t \big(V_1(s)+V_2(s)\big)\,ds, \qquad \forall f \in \mathbb{F},
$$

and for every $\ell \geq 2$,

$$
|\langle f, \Delta\nu_\ell(t)\rangle| \leq 3C(T)\int_0^t \big(V_{\ell-1}(s)+V_\ell(s)+V_{\ell+1}(s)\big)\,ds, \qquad \forall f \in \mathbb{F}.
$$

Taking the supremum over $f \in \mathbb{F}$ in the last two inequalities, for all $t \leq T$ we obtain

$$
V_\ell(t) \leq \begin{cases} 3C(T)\displaystyle\int_0^t \big(V_1(s)+V_2(s)\big)\,ds & \text{if } \ell = 1, \\ 3C(T)\displaystyle\int_0^t \big(V_{\ell-1}(s)+V_\ell(s)+V_{\ell+1}(s)\big)\,ds & \text{if } \ell \geq 2. \end{cases} \tag{3.17}
$$

Define $V(t) := \sum_{\ell=1}^\infty 2^{-\ell}V_\ell(t)$. Then (3.17) implies

$$
V(t) \leq 12C(T)\int_0^t V(s)\,ds. \tag{3.18}
$$

Also, since $|\langle f, \nu_\ell(t)\rangle| \leq 1$ for $f \in \mathbb{F}$, (3.1) implies $V_\ell(t) \leq 2$ for all $\ell \geq 1$, and hence $V(t) \leq 2$. Since $V(0) = 0$, an application of Gronwall's inequality then shows that $V(t) = 0$ for all $t \geq 0$, and hence, $V_\ell(t) = 0$ for all $t \geq 0$ and $\ell \geq 1$. In particular, this shows that

$$
\langle \mathbf{1}, \Delta\nu_k\rangle \equiv 0, \qquad k \geq 1. \tag{3.19}
$$

Moreover, by (3.16), $\Delta D_\ell \equiv 0$ for all $\ell \geq 2$. Taking the difference between equation (3.6), and the same equation, but with ν_ℓ, $D_{\ell+1}$ and η_ℓ replaced by $\tilde{\nu}_\ell$, $\tilde{D}_{\ell+1}$

and $\tilde{\eta}_\ell$, respectively, and using the identities $\Delta D_{\ell+1} \equiv 0$ and $\Delta \nu_\ell(0) = 0$, we see that for $\ell \geq 1$ and $f \in \mathbb{C}_b[0,\infty) \cap \mathbb{AC}_{\text{loc}}[0,\infty)$,

$$\langle f, \Delta\nu_\ell(t)\rangle = \int_0^t \langle f(\cdot + t - s)\vartheta^{t-s}(\cdot), \Delta\eta_\ell(s)\rangle\, ds. \tag{3.20}$$

To complete the proof, we use induction on ℓ to show that $\Delta\nu_\ell \equiv 0$ for $\ell \geq 1$. Let $\mathbb{AC}'[0, L) := \{f \in \mathbb{AC}[0, L) : ||f||_\infty \leq 1\}$. Since G has a density by Assumption II, $f(\cdot + t - s)\vartheta^{t-s}(\cdot) \in \mathbb{AC}'[0, L)$ for all $f \in \mathbb{AC}'[0, L)$ and $t, s \geq 0$. For $\ell = 1$, (3.2) and (3.19) with $k = 1$ imply $\langle f(\cdot + t - s)\vartheta^{t-s}, \Delta\eta_1(s)\rangle = 0$ and, therefore, $\Delta\nu_1 \equiv 0$ by (3.20). Furthermore, if $\Delta\nu_{\ell-1} \equiv 0$ for some $\ell \geq 2$, it follows from (3.4), (2.9) and (3.19), both with $k = \ell - 1$ and $k = \ell$, that

$$\begin{aligned}
&\left|\langle f(\cdot + t - s)\vartheta^{t-s}(\cdot), \Delta\eta_\ell(s)\rangle\right| \\
&\quad \leq \lambda \mathfrak{P}_d\big(\langle \mathbf{1}, \nu_{\ell-1}(s)\rangle, \langle \mathbf{1}, \nu_\ell(s)\rangle\big)\langle f(\cdot + t - s)\vartheta^{t-s}(\cdot), \Delta\nu_\ell(s)\rangle \\
&\quad \leq \lambda d \sup_{f \in \mathbb{AC}'[0,L)} \left|\langle f, \Delta\nu_\ell(s)\rangle\right|.
\end{aligned}$$

Together with (3.20) this implies

$$\sup_{f \in \mathbb{AC}'[0,L)} \left|\langle f, \Delta\nu_\ell(t)\rangle\right| \leq d \int_0^t \lambda \sup_{f \in \mathbb{AC}'[0,L)} \left|\langle f, \Delta\nu_\ell(s)\rangle\right| ds, \qquad \forall t \geq 0. \tag{3.21}$$

Since $\Delta\nu_\ell(0) = 0$, Gronwall's inequality shows $|\langle f, \Delta\nu_\ell(t)\rangle| = 0$ for $f \in \mathbb{AC}'[0, L)$ and hence, by linearity and a density argument, for $f \in \mathbb{C}_b[0, L)$. This proves $\Delta\nu_\ell(t) = 0$ for all $t \geq 0$. □

3.2. *A measure-valued PDE associated with the hydrodynamic limit.* The following is the main result of this section.

PROPOSITION 3.1. *Given $\nu(0) = (\nu_\ell(0); \ell \geq 1) \in \mathbb{S}$ and $\lambda > 0$, suppose $\nu = (\nu_\ell)_{\ell\geq1} \in \mathbb{D}_\mathbb{S}[0,\infty)$ satisfies the following*: (2.10) *holds and for every $\ell \geq 1$ and $t \geq 0$,* (2.11) *holds, with D_ℓ and η_ℓ defined as in* (2.12) *and* (2.14), *respectively, and for $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$,*

$$\begin{aligned}
\langle \varphi(\cdot, t), \nu_\ell(t)\rangle &= \langle \varphi(\cdot, 0), \nu_\ell(0)\rangle + \int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s) - \varphi(\cdot, s)h(\cdot), \nu_\ell(s)\rangle\, ds \\
&\quad + \int_0^t \varphi(0, s)\, dD_{\ell+1}(s) + \int_0^t \langle \varphi(\cdot, s), \eta_\ell(s)\rangle\, ds.
\end{aligned} \tag{3.22}$$

Then ν is a solution to the hydrodynamic equations associated with $(\lambda, \nu(0))$.

The proof of Proposition 3.1 relies on the following lemma. Denote by $\tilde{\mathbb{M}}$ the space of Radon measures on $\mathbb{R}^2$ whose supports lie in $[0, L) \times \mathbb{R}_+$, and denote the integral with respect to any Radon measure Θ on $\mathbb{R}^2$ by

$$\Theta(\varphi) = \int_{\mathbb{R}^2} \varphi(x, s)\Theta(dx\, ds), \qquad \varphi \in \mathbb{C}_c\big([0, L) \times \mathbb{R}_+\big).$$

LEMMA 3.2. *Given a Radon measure* $\Theta \in \tilde{\mathbb{M}}$, *suppose* $\mu = \{\mu(t); t \geq 0\}$ *in* $\mathbb{D}_{\mathbb{M}_F[0,L)}[0,\infty)$ *satisfies* $\int_0^t \langle h, \mu(s)\rangle\, ds < \infty$, *and and for every* $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$,

$$-\int_0^\infty \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s) - \varphi(\cdot,s)h(\cdot), \mu(s)\rangle\, ds = \Theta(\varphi). \tag{3.23}$$

Then, for every $f \in \mathbb{C}_c[0,L)$ *and* $t \geq 0$,

$$\langle f, \mu(t)\rangle = \int_{[0,L)\times[0,t]} f(x+t-s)\frac{\overline{G}(x+t-s)}{\overline{G}(x)}\Theta(dx\, ds). \tag{3.24}$$

Equation (3.23) is called the *abstract age equation* and was studied extensively in Section 4.3. of [24] Lemma 3.2 essentially follows from Corollary 4.17 and equations (4.24), (4.45), (4.46) and (4.55) in [24]; for a detailed proof, see Lemma 3.2 of [5].

PROOF OF PROPOSITION 3.1. Clearly, we only need to show that (2.13) holds for all $t \geq 0$, $\ell \geq 1$ and $f \in \mathbb{C}_b[0,L)$. Fix $\ell \geq 1$, and consider the linear functional ξ_ℓ on $\mathbb{C}_c(\mathbb{R}^2)$ defined by

$$\begin{aligned}\xi_\ell(\varphi) &:= \langle \varphi(\cdot,0), \nu_\ell(0)\rangle + \int_{[0,\infty)} \varphi(0,s)\, dD_{\ell+1}(s) \\ &\quad + \int_0^\infty \langle \varphi(\cdot,s), \eta_\ell(s)\rangle\, ds.\end{aligned} \tag{3.25}$$

By (2.14) and (2.9), for $m \in [0,L)$, $T \in [0,\infty)$ and every $\varphi \in \mathbb{C}_c(\mathbb{R}^2)$ with $\operatorname{supp}(\varphi) \subset [0,m] \times [0,T]$,

$$\left|\langle \varphi(\cdot,s), \eta_\ell(s)\rangle\right| \leq \|\varphi\|_\infty \lambda d. \tag{3.26}$$

Hence, since $D_{\ell+1}$ is nondecreasing,

$$\left|\xi_\ell(\varphi)\right| \leq \|\varphi\|_\infty\left(\left|\nu_\ell(0)\right|_{TV} + D_{\ell+1}(T) + C(T)\right), \tag{3.27}$$

with $C(T) := \lambda d T < \infty$, and $D_{\ell+1}(T) < \infty$ by (2.10). Moreover, $\xi_\ell(\varphi) = 0$ for all φ such that $\operatorname{supp}(\varphi) \cap [0,L) \times \mathbb{R}_+ = \varnothing$. Hence, ξ is a Radon measure on $\mathbb{R}^2$ with support in $[0,L) \times \mathbb{R}_+$, that is, $\xi \in \tilde{\mathbb{M}}$. Moreover, since φ has compact support, the left-hand side of (3.22) is equal to zero for sufficiently large t. Therefore, sending $t \to \infty$ in (3.22), we have for all $\varphi \in \mathbb{C}_c^{1,1}([0,L) \times \mathbb{R}_+)$,

$$-\int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s) - \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle\, ds = \xi_\ell(\varphi). \tag{3.28}$$

Thus, ν_ℓ satisfies the abstract age equation associated with $\xi_\ell \in \tilde{\mathbb{M}}$ and h. Therefore, by Lemma 3.2 and (3.25), for all $f \in \mathbb{C}_c[0,L)$ and $t \geq 0$,

$$\begin{aligned}\langle f, \nu_\ell(t)\rangle &= \int_{[0,L)\times[0,t]} f(x+t-s)\frac{\overline{G}(x+t-s)}{\overline{G}(x)}\xi_\ell(dx\, ds) \\ &= \left\langle f(\cdot+t)\frac{\overline{G}(\cdot+t)}{\overline{G}(\cdot)}, \nu_\ell(0)\right\rangle\end{aligned}$$

$$
\begin{aligned}
&+ \int_{[0,t]} f(t-s)\overline{G}(t-s)\, dD_{\ell+1}(s) \\
&+ \int_0^t \left\langle f(\cdot + t - s)\frac{\overline{G}(\cdot + t - s)}{\overline{G}(\cdot)}, \eta_\ell(s)\right\rangle ds.
\end{aligned} \tag{3.29}
$$

Since for $t \geq 0$, the right-hand side of (3.29) is finite for every $f \in \mathbb{C}_b[0, L)$, the relation (3.29) can be extended to all $f \in \mathbb{C}_b[0, L)$ by an application of the dominated convergence theorem, and (2.13) follows. This completes the proof. □

4. State dynamics. In Section 4.1, we express the state descriptor $\nu^{(N)}$ in terms of primitives of the networks. This will be required to justify the existence of compensators of certain processes in Section 5. We also introduce some auxiliary processes in Section 4.2, and use them to derive dynamical equations for the state variables in Section 4.3.

REMARK 4.1. To make it easy to follow the notation, throughout the paper, we use the superscript i to denote queue indices and subscript j to denote jobs.

4.1. *State variables.* For each job j, let $\gamma_j^{(N)}$, $\alpha_j^{(N)}$ and $\beta_j^{(N)}$, respectively, represent the time at which job j arrives into the system, enters service and departs the queue on completing service. Note that $\gamma_j^{(N)} = \alpha_j^{(N)}$ if the job is routed to an empty queue, and $\beta_j^{(N)} = \alpha_j^{(N)} + v_j$ where $\{v_j\}$ is the i.i.d. sequence of service times. We use the convention that jobs initially in the network are indexed by nonpositive numbers $j = j_0, \dots, 0$ with $j_0 := -X^{(N)}(0) + 1$ being the smallest job index, where recall that $X^{(N)}(t)$ represents the total number of jobs in system at time t. We also assume that jobs that entered service earlier get smaller indices. Jobs that arrive after time 0 are given indices $j \geq 1$ in the order of their arrival time ($0 < \gamma_j^{(N)} < \gamma_{j+1}^{(N)}$ for $j \geq 1$, almost surely). Then the age $a_j^{(N)}(t)$ of job j at time t takes the form

$$
a_j^{(N)}(t) := \begin{cases} 0 & \text{if } t < \alpha_j^{(N)}, \\ t - \alpha_j^{(N)} & \text{if } \alpha_j^{(N)} \leq t < \beta_j^{(N)}, \\ v_j & \text{if } t \geq \beta_j^{(N)}. \end{cases} \tag{4.1}
$$

We also assign to each queue an index $i \in \{1, \dots, N\}$. To implement the SQ(d) routing algorithm, upon arrival, each job $j \geq 1$ chooses a vector $\iota_j = (\iota_j(1), \dots, \iota_j(\mathrm{d}))$ of d indices, each chosen independently and uniformly at random from the set $\{1, \dots, N\}$ (in practice, it would be more natural to sample d queues at random without replacement, but we choose the former routing procedure for simplicity; the effect of the difference vanishes in the hydrodynamic limit). The job is then routed to the queue with the shortest length amongst the chosen indices, where if there are multiple queues of minimal length, then one of them is chosen uniformly at random. We denote the index of the queue to which job j is routed by $\kappa_j^{(N)}$:

$$\kappa_j^{(N)} \sim \operatorname{Unif}\bigl(\operatorname{argmin}\{X^{(N),\iota_j(1)}(\gamma_j^{(N)}-),\dots,X^{(N),\iota_j(d)}(\gamma_j^{(N)}-)\}\bigr), \tag{4.2}$$

where recall $X^{(N),i}(t)$ is the number of jobs in the ith queue at time t. With a slight abuse of notation, we also use $\kappa_j^{(N)}(t) := \mathbb{1}_{\{t\geq\gamma_j^{(N)}\}}\kappa_j^{(N)}$ to denote the queue index process.

In our Markovian description, the initial state of the network is completely determined by $R_E^{(N)}(0)$ from (2.2) and $\nu^{(N)}(0)$. Since the routing allocation is symmetric with respect to queues, if the vector of initial queue lengths and ages is symmetric, the vector of queue lengths at any time t is also exchangeable. Thus, the empirical measure $\nu^{(N)}$ captures the essential features of the dynamics. However, it will prove convenient to also refer to a more detailed description, in which $\kappa_j^{(N)}$ is specified for each job initially in the system. According to our indexing convention, the $X^{(N)}(0)$ job initially in the network have indices in the set $\{j_0 := -X^{(N)}(0)+1,\dots,0\}$, and $\langle\mathbf{1},\nu_1(0)\rangle$ jobs initially in service have indices in the set $\{j_0,\dots,j_0+\langle\mathbf{1},\nu_1^{(N)}(0)\rangle-1\}$. Now, let

$$I_0 := \bigl(R_E^{(N)}(0), a_j^{(N)}(0), \kappa_j^{(N)}; j=-X^{(N)}(0)+1,\dots,0\bigr). \tag{4.3}$$

Assumption III(a) can be expressed in terms of the above notation as follows: for every finite subset $\mathcal{K}\subset\{-X^{(N)}(0)+1,\dots,\dots,0\}$ of jobs initially in system,

$$\mathbb{P}\{v_j > b_j; j\in\mathcal{K}|I_0\} = \prod_{j\in\mathcal{K}}\frac{\overline{G}(a_j^{(N)}(0)+b_j)}{\overline{G}(a_j^{(N)}(0))}, \qquad b_j\geq 0. \tag{4.4}$$

In other words, for every job j not initially in service, v_j is independent of I_0.

We now express the measures $\nu_\ell^{(N)}$ in terms of the primitives defined above. A job j receives service during the interval $[\alpha_j^{(N)},\beta_j^{(N)})$, and hence,

$$\mathcal{V}^{(N)}(t) := \{j\geq j_0 : \alpha_j^{(N)}\leq t<\beta_j^{(N)}\} \tag{4.5}$$

is the set of indices of jobs receiving service at time t. Also, for $t\in[\gamma_j^{(N)},\infty)$ let $\chi_j^{(N)}(t)$ denote the length at time t of the queue to which job j was routed. In other words,

$$\chi_j^{(N)}(t) = X^{(N),\kappa_j^{(N)}}(t), \qquad t\geq\gamma_j^{(N)}. \tag{4.6}$$

The value of $\chi_j^{(N)}(t)$ for $t<\gamma_j^{(N)}$ is irrelevant. Therefore, for $\ell\geq 1$,

$$\mathcal{U}_\ell^{(N)}(t) := \{j\geq j_0 : \mathbb{1}_{\{\gamma_j^{(N)}\leq t\}}\chi_j^{(N)}(t)\geq\ell\} \tag{4.7}$$

is the set of jobs in queues with length at least ℓ at time t. Using this notation, we can write

$$\nu_\ell^{(N)}(t) := \sum_{j=j_0}^{\infty} \mathbb{1}_{\{\alpha_j^{(N)} \le t\}} \mathbb{1}_{\{\beta_j^{(N)} > t\}} \mathbb{1}_{\{\chi_j^{(N)}(t) \ge \ell\}} \delta_{a_j^{(N)}(t)} \tag{4.8}$$

$$= \sum_{j=j_0}^{\infty} \mathbb{1}_{\{j \in \mathcal{V}^{(N)}(t)\}} \mathbb{1}_{\{j \in \mathcal{U}_\ell^{(N)}(t)\}} \delta_{a_j^{(N)}(t)}. \tag{4.9}$$

The pair $(R_E^{(N)}, \nu^{(N)})$ with $\nu^{(N)} = (\nu_\ell^{(N)}; \ell \ge 1)$ is the state descriptor of the N-server network.

4.2. *Auxiliary processes and filtration.* To describe the dynamics of $\nu^{(N)}$, it will be convenient to introduce a number of auxiliary processes. For every queue $i \in \{1, \dots, N\}$, let $E^{(N),i}$ denote the cumulative arrival process to queue i, defined as

$$E^{(N),i}(t) := \sum_{j=1}^{\infty} \mathbb{1}_{\{\gamma_j^{(N)} \le t\}} \mathbb{1}_{\{\kappa_j^{(N)} = i\}}, \qquad t \ge 0. \tag{4.10}$$

For $\ell \ge 1$, let $\nu_\ell^{(N),i}(t)$ denote the measure that has a Dirac delta mass at the age of the job in service at queue i at time t if that queue has length at least ℓ: that is, for $t \ge 0$,

$$\nu_\ell^{(N),i}(t) := \sum_{j=j_0}^{\infty} \mathbb{1}_{\{\alpha_j^{(N)} \le t\}} \mathbb{1}_{\{\beta_j^{(N)} > t\}} \mathbb{1}_{\{\chi_j^{(N)}(t) \ge \ell\}} \mathbb{1}_{\{\kappa_j^{(N)} = i\}} \delta_{a_j^{(N)}(t)}. \tag{4.11}$$

Note that $\nu_\ell^{(N),i}(t)$ always has mass either zero or 1, and clearly,

$$\nu_\ell^{(N)} = \sum_{i=1}^{N} \nu_\ell^{(N),i}. \tag{4.12}$$

Fix $\ell \ge 1$. For $\varphi \in \mathbb{C}_b([0, L) \times \mathbb{R}_+)$, let $\mathcal{R}_{\varphi,\ell}^{(N)}$ be the cumulative φ-weighted routing measure process to queues with length exactly $\ell - 1$, defined as follows for all $t \ge 0$:

$$\mathcal{R}_{\varphi,\ell}^{(N)}(t) := \begin{cases} \displaystyle\sum_{i=1}^{N} \int_{(0,t]} \varphi(0, s)\big(1 - \langle \mathbf{1}, \nu_1^{(N),i}(s-)\rangle\big)\, dE^{(N),i}(s) \\ \quad \text{if } \ell = 1, \\ \displaystyle\sum_{i=1}^{N} \int_{(0,t]} \langle \varphi(\cdot, s), \nu_{\ell-1}^{(N),i}(s-) - \nu_\ell^{(N),i}(s-)\rangle\, dE^{(N),i}(s), \\ \quad \text{if } \ell \ge 2. \end{cases} \tag{4.13}$$

Roughly speaking, $\mathcal{R}^{(N)}_{\varphi,\ell}(t)$ captures the cumulative effect on the measure $\nu^{(N)}_\ell$ due to jobs routed in the interval $[0,t]$. Indeed, in both cases, $\varphi(\cdot,s)$ in the integral is evaluated at the age of the job in service at queue i. Note that when $\ell=1$, since $\langle \mathbf{1}, \nu^{(N),i}(s-)\rangle = X^{(N),i}(s-)$, we can also write

$$\mathcal{R}^{(N)}_{\varphi,1}(t) = \sum_{i=1}^{N} \int_{(0,t]} \varphi(0,s) \mathbb{1}_{\{X^{(N),i}(s-)=0\}}\, dE^{(N),i}(s). \tag{4.14}$$

Next, we turn to the counting process $D^{(N)}_\ell = \{D^{(N)}_\ell(t); t\geq 0\}$ of departures from queues with length at least ℓ right before departure. For conciseness, we use the following notation for values of the queue length of job j right after its arrival time or service entry and right before its departure time:

$$\begin{aligned} \chi_j^{E,(N)} &:= \chi_j^{(N)}(\gamma_j^{(N)}), \\ \chi_j^{K,(N)} &:= \chi_j^{(N)}(\alpha_j^{(N)}), \\ \chi_j^{D,(N)} &:= \chi_j^{(N)}(\beta_j^{(N)}-), \end{aligned} \tag{4.15}$$

where $\chi_j^{(N)}(\cdot)$ is the queue length process defined in (4.6). Then we have

$$D^{(N)}_\ell(t) = \sum_{j=j_0}^{\infty} \mathbb{1}_{\{\beta_j^{(N)}\leq t\}} \mathbb{1}_{\{\chi_j^{D,(N)}\geq \ell\}}, \qquad t\geq 0. \tag{4.16}$$

Note that $D^{(N)} := D^{(N)}_1$ is the total cumulative departure process.

REMARK 4.2. Since a queue is never empty just prior to a departure or right after a service entry, we have $\chi_j^{D,(N)}\geq 1$ and $\chi_j^{K,(N)}\geq 1$. Also, a simple mass balance shows that

$$D^{(N)}(t) + \langle \mathbf{1}, \nu_1^{(N)}(t)\rangle \leq X^{(N)}(0) + E^{(N)}(t). \tag{4.17}$$

For $\varphi \in \mathbb{C}_b([0,L)\times \mathbb{R}_+)$, let $\mathcal{D}^{(N)}_{\varphi,\ell}$ be the cumulative φ-weighted departure process from queues of length at least ℓ, defined by

$$\mathcal{D}^{(N)}_{\varphi,\ell}(t) := \sum_{j=j_0}^{\infty} \varphi\big(v_j, \beta_j^{(N)}\big) \mathbb{1}_{\{\beta_j^{(N)}\leq t\}} \mathbb{1}_{\{\chi_j^{D,(N)}\geq \ell\}}, \qquad t\geq 0. \tag{4.18}$$

Clearly, $\mathcal{D}^{(N)}_{\mathbf{1},\ell} = D^{(N)}_\ell$, and hence by (4.17),

$$\big|\mathcal{D}^{(N)}_{\varphi,\ell}(t)\big| \leq \|\varphi\|_\infty \big(X^{(N)}(0) + E^{(N)}(t)\big), \qquad t\geq 0. \tag{4.19}$$

For $i \in \{1, \dots, N\}$, let $D^{(N),i}$ denote the departure process from queue i. Then

$$D^{(N),i}(t) = \sum_{j=j_0}^{\infty} \mathbb{1}_{\{\beta_j^{(N)} \le t\}} \mathbb{1}_{\{\kappa_j^{(N)} = i\}}, \qquad t \ge 0, \tag{4.20}$$

$$X^{(N),i}(t) = X^{(N),i}(0) + E^{(N),i}(t) - D^{(N),i}(t), \qquad t \ge 0. \tag{4.21}$$

Finally, we define the filtration $\{\mathcal{F}_t^{(N)}; t \ge 0\}$ generated by the initial conditions of the network, see (4.3), plus the filtrations $\{\mathcal{F}_t^{E^{(N),i}}\}$ and $\{\mathcal{F}_t^{D^{(N),i}}\}$ generated by $E^{(N),i}$ and $D^{(N),i}$, respectively, $i = 1, \dots, N$. In other words,

$$\mathcal{F}_t^{(N)} := \bigvee_{i=1}^{N} (\mathcal{F}_t^{E^{(N),i}} \vee \mathcal{F}_t^{D^{(N),i}}) \vee \sigma(I_0). \tag{4.22}$$

It is easy to see that all state variables and auxiliary processes are $\{\mathcal{F}_t^{(N)}\}$-adapted.

REMARK 4.3. It is possible to show that $\{\mathcal{F}_t^{(N)}\}$ is also equal to the filtration generated by the age and queue index processes $a_j^{(N)}(\cdot), \kappa_j^{(N)}(\cdot)$; $j \ge 1$. However, our definition allows us to exploit results from [12] in Section 5 to identify compensators of certain processes.

REMARK 4.4. One can also show that $\{(R_E^{(N)}(t), \nu^{(N)}(t)); t \ge 0\}$ is a Markov process with respect to the filtration $\{\mathcal{F}_t^{(N)}; t \ge 0\}$; but, since we do not use this property, we do not prove it.

4.3. *Equations governing the dynamics of the N-server network.* We now describe the dynamics of the state descriptor $\nu^{(N)}$ for a fixed N. Our main result (Proposition 4.5) does not require all our assumptions on the arrival process and service time distribution, but instead holds for a very general class of networks and load balancing algorithms, as long as all arrival and departure times are distinct, almost surely. To make this notion precise, denote by $\Omega_{s,\delta}$ the set of realizations for which at most one arrival or one departure occurs during $(s, s+\delta]$. Also, define Ω_t to be the set of realizations for which there exists a partition $\{(\frac{k}{n}, \frac{k+1}{n}]; k = 0, \dots, \lfloor nt \rfloor\}$ of $(0, t]$ such that at most one arrival or one departure occurs in each subinterval, that is,

$$\Omega_t := \bigcup_{n=1}^{\infty} \bigcap_{k=0}^{\lfloor nt \rfloor} \Omega_{\frac{k}{n}, \frac{1}{n}}. \tag{4.23}$$

Proposition 4.5 below establishes an implicit relation between $\nu_\ell^N(t)$ and the departure and routing processes. This result is not specific to the $SQ(d)$ algorithm, and in fact, we believe that it holds true for any arbitrary algorithm that routes jobs immediately upon arrival, should the appropriate arrival processes $E^{(N),i}$ corresponding to the algorithm be substituted in (4.13).

PROPOSITION 4.5. *Consider an N-server network with any arrival process, service times and load balancing algorithm such that $\mathbb{P}\{\Omega_t\} = 1$ for every $t \geq 0$. Then, for $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, almost surely, for $\ell \geq 1$ and $t \geq 0$,*

$$\begin{aligned}\langle \varphi(\cdot, t), \nu_\ell^{(N)}(t)\rangle = {}& \langle \varphi(\cdot, 0), \nu_\ell^{(N)}(0)\rangle \\ & + \int_0^t \langle \varphi_s(\cdot, s) + \varphi_x(\cdot, s), \nu_\ell^{(N)}(s)\rangle \, ds - \mathcal{D}_{\varphi,\ell}^{(N)}(t) \\ & + \int_{[0,t]} \varphi(0, s) \, dD_{\ell+1}^{(N)}(s) + \mathcal{R}_{\varphi,\ell}^{(N)}(t)\end{aligned} \tag{4.24}$$

and

$$\langle \mathbf{1}, \nu_\ell^{(N)}(t)\rangle = \langle \mathbf{1}, \nu_\ell^{(N)}(0)\rangle - D_\ell^{(N)}(t) + D_{\ell+1}^{(N)}(t) + \mathcal{R}_{\mathbf{1},\ell}^{(N)}(t). \tag{4.25}$$

The rest of this section is devoted to the proof of this proposition. Throughout this section, for ease of notation, $\sum_j$ is used to denote the sum over all job indices $j \in \{j_0, \ldots, -1, 0, 1, \ldots\}$.

Fix $\ell \geq 1$, $t \geq 0$. For $\varphi \in \mathbb{C}_b([0, L) \times [0, \infty))$, since $s \mapsto \langle \varphi(\cdot, s), \nu_\ell^{(N)}(s)\rangle$ is right-continuous, we can write

$$\begin{aligned}& \langle \varphi(\cdot, t), \nu_\ell^{(N)}(t)\rangle - \langle \varphi(\cdot, 0), \nu_\ell^{(N)}(0)\rangle \\ & \quad = \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \left[\left\langle \varphi\left(\cdot, \frac{k+1}{n}\right), \nu_\ell^{(N)}\left(\frac{k+1}{n}\right)\right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_\ell^{(N)}\left(\frac{k}{n}\right)\right\rangle \right] \\ & \quad = \lim_{n\to\infty} \left(\mathcal{I}_1^{(n)} + \mathcal{I}_2^{(n)}\right),\end{aligned} \tag{4.26}$$

where

$$\mathcal{I}_1^{(n)}(t) := \sum_{k=0}^{\lfloor nt \rfloor} \left\langle \varphi\left(\cdot, \frac{k+1}{n}\right) - \varphi\left(\cdot, \frac{k}{n}\right), \nu_\ell^{(N)}\left(\frac{k+1}{n}\right)\right\rangle \tag{4.27}$$

and

$$\mathcal{I}_2^{(n)}(t) := \sum_{k=0}^{\lfloor nt \rfloor} \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_\ell^{(N)}\left(\frac{k+1}{n}\right)\right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_\ell^{(N)}\left(\frac{k}{n}\right)\right\rangle. \tag{4.28}$$

By Riemann integrability (see page 21 of [5] for full details), we have

$$\lim_{n\to\infty} \mathcal{I}_1^{(n)}(t) = \int_0^t \langle \varphi_s(\cdot, s), \nu_\ell^{(N)}(s)\rangle \, ds. \tag{4.29}$$

To compute the limit of $\mathcal{I}_2^{(n)}(t)$, first fix $s, \delta \geq 0$ and use the expressions for $\nu_\ell^{(N)}$, $a_j^{(N)}$ $\mathcal{V}^{(N)}$ and $\mathcal{U}_\ell^{(N)}$ in (4.9), (4.1), (4.5) and (4.7) to write

$$\begin{aligned}\nu_\ell^{(N)}(s+\delta) &= \sum_j \mathbb{1}_{\{j \in \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s+\delta)\}} \delta_{a_j^{(N)}(s+\delta)} \\ &= \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3,\end{aligned} \tag{4.30}$$

where

$$\mathcal{J}_1 := \sum_j \mathbb{1}_{\{j\in\mathcal{V}^{(N)}(s+\delta)\cap\mathcal{U}_\ell^{(N)}(s+\delta)\setminus\mathcal{V}^{(N)}(s)\}}\delta_{a_j^{(N)}(s+\delta)},$$

$$\mathcal{J}_2 := \sum_j \mathbb{1}_{\{j\in\mathcal{V}^{(N)}(s+\delta)\cap\mathcal{V}^{(N)}(s)\cap\mathcal{U}_\ell^{(N)}(s+\delta)\setminus\mathcal{U}_\ell^{(N)}(s)\}}\delta_{a_j^{(N)}(s)+\delta},$$

$$\mathcal{J}_3 := \sum_j \mathbb{1}_{\{j\in\mathcal{V}^{(N)}(s+\delta)\cap\mathcal{V}^{(N)}(s)\cap\mathcal{U}_\ell^{(N)}(s+\delta)\cap\mathcal{U}_\ell^{(N)}(s)\}}\delta_{a_j^{(N)}(s)+\delta}.$$

Next, applying the identity $C\cap F\cap\tilde{F} = \tilde{F}\setminus\{[\tilde{F}\setminus C]\cup[(\tilde{F}\cap C)\setminus F]\}$ with $C = \mathcal{V}^{(N)}(s+\delta)$, $F = \mathcal{U}_\ell^{(N)}(s+\delta)$ and $\tilde{F} = \mathcal{V}^{(N)}(s)\cap\mathcal{U}_\ell^{(N)}(s)$, we have

$$\begin{aligned}\mathcal{J}_3 = {} & \sum_j \mathbb{1}_{\{j\in\mathcal{V}^{(N)}(s)\cap\mathcal{U}_\ell^{(N)}(s)\}}\delta_{a_j^{(N)}(s)+\delta} - \mathcal{J}_3' \\ & - \sum_j \mathbb{1}_{\{j\in[\mathcal{V}^{(N)}(s)\cap\mathcal{U}_\ell^{(N)}(s)\cap\mathcal{V}^{(N)}(s+\delta)]\setminus\mathcal{U}_\ell^{(N)}(s+\delta)\}}\delta_{a_j^{(N)}(s)+\delta},\end{aligned} \tag{4.31}$$

where

$$\mathcal{J}_3' := \sum_j \mathbb{1}_{\{j\in[\mathcal{V}^{(N)}(s)\cap\mathcal{U}_\ell^{(N)}(s)]\setminus\mathcal{V}^{(N)}(s+\delta)\}}\delta_{a_j^{(N)}(s)+\delta}.$$

Note we have not used any property of $\mathcal{U}_\ell^{(N)}$ in this calculation. We now restrict to realizations $\Omega_{s,\delta}$ when there is only a single arrival or departure during $(s, s+\delta]$.

LEMMA 4.6. *For $\ell \geq 1$ and $\varphi \in \mathbb{C}_b([0,L)\times[0,\infty))$, $s \geq 0$ and $\delta > 0$, on $\Omega_{s,\delta}$,*

$$\begin{aligned}&\langle\varphi(\cdot,s), \nu_\ell^{(N)}(s+\delta)\rangle \\ &\quad = \langle\varphi(\cdot+\delta,s), \nu_\ell^{(N)}(s)\rangle \\ &\qquad - \sum_j \varphi\big(a_j^{(N)}(s)+\delta, s\big)\mathbb{1}_{\{\beta_j^{(N)}\in(s,s+\delta]\}}\mathbb{1}_{\{\chi_j^{D,(N)}\geq\ell\}} \\ &\qquad + \sum_j \varphi\big(a_j^{(N)}(s+\delta), s\big)\mathbb{1}_{\{\alpha_j^{(N)}\in(s,s+\delta]\}}\mathbb{1}_{\{\chi_j^{K,(N)}\geq\ell\}} \\ &\qquad + \mathbb{1}_{\{\ell\geq2\}}\sum_j \varphi\big(a_j^{(N)}(s+\delta), s\big)\mathbb{1}_{\{\alpha_j^{(N)}\leq s\}}\mathbb{1}_{\{\beta_j^{(N)}>s\}} \\ &\qquad \times \mathbb{1}_{\{\chi_j^{(N)}(s)=\ell-1\}}\sum_{j'\geq1}\mathbb{1}_{\{\gamma_{j'}^{(N)}\in(s,s+\delta]\}}\mathbb{1}_{\{\kappa_{j'}^{(N)}=\kappa_j^{(N)}\}}.\end{aligned} \tag{4.32}$$

REMARK 4.7. Since $\chi_j^{D,(N)} \geq 1$ and $\chi_j^{K,(N)} \geq 1$, for $\ell = 1$ (4.32) reduces to

$$\begin{aligned}\langle \varphi(\cdot, s), \nu_1^{(N)}(s+\delta)\rangle = {} & \langle \varphi(\cdot + \delta, s), \nu_1^{(N)}(s)\rangle \\ & - \sum_j \varphi(a_j^{(N)}(s) + \delta, s)\mathbb{1}_{\{\beta_j^{(N)} \in (s, s+\delta]\}} \\ & + \sum_j \varphi(a_j^{(N)}(s+\delta), s)\mathbb{1}_{\{\alpha_j^{(N)} \in (s, s+\delta]\}}. \end{aligned} \tag{4.33}$$

PROOF OF LEMMA 4.6. We first simplify the terms $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ in (4.30). A job j receives service at time $s+\delta$, but not at s if and only if it entered service during $(s, s+\delta]$. Moreover, on $\Omega_{s,\delta}$, there could have been no arrivals in $(s, s+\delta]$ and so, the length of the job's queue is constant from the service entry time to $s+\delta$. This implies $j \in \mathcal{U}_\ell^{(N)}(s+\delta)$ if and only if $\chi_j^{K,(N)} \geq \ell$. Thus,

$$\mathcal{J}_1 = \sum_j \mathbb{1}_{\{\alpha_j^{(N)} \in (s, s+\delta]\}} \mathbb{1}_{\{\chi_j^{K,(N)} \geq \ell\}} \delta_{a_j^{(N)}(s+\delta)}. \tag{4.34}$$

We now analyze the term $\mathcal{J}_2$. If a job j received service throughout the period $(s, s+\delta]$, that is $j \in \mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta)$, then the corresponding queue could not have been empty at time s, and on $\Omega_{s,\delta}$, the difference between the queue length at time $s+\delta$ and time s is either zero (if there were no arrivals to that queue) or one (if there was precisely one arrival to that queue). Therefore, when $\ell = 1$,

$$\mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_1^{(N)}(s+\delta) \backslash \mathcal{U}_1^{(N)}(s) = \varnothing, \tag{4.35}$$

and $\mathcal{J}_2 = 0$, whereas for $\ell = 2$, using the representation (4.10) for $E^{(N),i}$,

$$\begin{aligned}\mathcal{J}_2 &= \sum_j \mathbb{1}_{\{\alpha_j^{(N)} \leq s\}} \mathbb{1}_{\{\beta_j^{(N)} > s\}} \mathbb{1}_{\{\chi_j^{(N)}(s) = \ell - 1\}} E^{(N),\kappa_j^{(N)}}(s, s+\delta] \delta_{a_j^{(N)}(s+\delta)} \\ &= \sum_j \mathbb{1}_{\{\alpha_j^{(N)} \leq s\}} \mathbb{1}_{\{\beta_j^{(N)} > s\}} \mathbb{1}_{\{\chi_j^{(N)}(s) = \ell - 1\}} \delta_{a_j^{(N)}(s+\delta)} \sum_{j' \geq 1} \mathbb{1}_{\{\gamma_{j'}^{(N)} \in (s, s+\delta], \kappa_{j'}^{(N)} = \kappa_j^{(N)}\}}.\end{aligned}$$

For the third term $\mathcal{J}_3$, we use (4.31). First, note that by the form (4.9) of $\nu_\ell^{(N)}$,

$$\left\langle \varphi(\cdot, s), \sum_j \mathbb{1}_{\{j \in \mathcal{V}^{(N)}(s) \cap \mathcal{U}_\ell^{(N)}(s)\}} \delta_{a_j^{(N)}(s)+\delta} \right\rangle = \langle \varphi(\cdot + \delta), s), \nu_\ell^{(N)}(s)\rangle. \tag{4.36}$$

Next, note that a job j departed a queue during $(s, s+\delta]$ if and only if $j \in \mathcal{V}^{(N)}(s) \backslash \mathcal{V}^{(N)}(s+\delta)$. Moreover, on $\Omega_{s,\delta}$ there were no arrivals during $(s, s+\delta]$, and hence, the queue length was constant on $(s, \beta_j^{(N)}-)$. Therefore, $j \in \mathcal{U}_\ell^{(N)}(s)$ if and only if $\chi_j^{D,(N)} \geq \ell$. Hence,

$$\mathcal{J}_3' = \sum_j \mathbb{1}_{\{\beta_j^{(N)} \in (s, s+\delta]\}} \mathbb{1}_{\{\chi_j^{D,(N)} \geq \ell\}} \delta_{a_j^{(N)}(s)+\delta}. \tag{4.37}$$

Finally, for the last term on the right-hand side of (4.31), note that if a job j receives service at a queue during $(s, s+\delta]$, then that queue length is nondecreasing on that interval. Therefore,

$$\bigl[\mathcal{V}^{(N)}(s) \cap \mathcal{V}^{(N)}(s+\delta) \cap \mathcal{U}_\ell^{(N)}(s)\bigr] \backslash \mathcal{U}_\ell^{(N)}(s+\delta) = \varnothing. \tag{4.38}$$

The result follows from (4.30), (4.31) and (4.34)–(4.38). □

We continue with the identification of the limit of $\mathcal{I}_2^{(n)}(t)$. Since $\mathbb{P}\{\Omega_t\} = 1$ by assumption, there exists $n_0 \in \mathbb{N}$ such that almost surely, the identity (4.32) holds with $\delta = 1/n$ and $s = k/n$ simultaneously for every $n \geq n_0$ and $k = 0, 1, \dots, \lfloor nt \rfloor$. Substituting (4.32) with $\delta = 1/n$ and $s = k/n$ into (4.28), we have almost surely

$$\begin{aligned}
\mathcal{I}_2^{(n)}(t) = {} & \sum_{k=0}^{\lfloor nt \rfloor} \Bigl[\Bigl\langle \varphi\Bigl(\cdot + \frac{1}{n}, \frac{k}{n}\Bigr) - \varphi\Bigl(\cdot, \frac{k}{n}\Bigr), \nu_\ell^{(N)}\Bigl(\frac{k}{n}\Bigr)\Bigr\rangle\Bigr] \\
& - \sum_{k=0}^{\lfloor nt \rfloor} \sum_j \varphi\Bigl(a_j^{(N)}\Bigl(\frac{k}{n}\Bigr) + \frac{1}{n}, \frac{k}{n}\Bigr) \mathbb{1}_{\{\beta_j^{(N)} \in (\frac{k}{n}, \frac{k+1}{n}]\}} \mathbb{1}_{\{\chi_j^{D,(N)} \geq \ell\}} \\
& + \sum_{k=0}^{\lfloor nt \rfloor} \sum_j \varphi\Bigl(a_j^{(N)}\Bigl(\frac{k+1}{n}\Bigr), \frac{k}{n}\Bigr) \mathbb{1}_{\{\alpha_j^{(N)} \in (\frac{k}{n}, \frac{k+1}{n}]\}} \mathbb{1}_{\{\chi_j^{K,(N)} \geq \ell\}} \\
& + \mathbb{1}_{\{\ell \geq 2\}} \sum_{k=0}^{\lfloor nt \rfloor} \sum_j \sum_{j' \geq 1} \varphi\Bigl(a_j^{(N)}\Bigl(\frac{k+1}{n}\Bigr), \frac{k}{n}\Bigr) \mathbb{1}_{\{\alpha_j^{(N)} \leq \frac{k}{n}\}} \mathbb{1}_{\{\beta_j^{(N)} > \frac{k}{n}\}} \\
& \times \mathbb{1}_{\{\chi_j^{(N)}(\frac{k}{n}) = \ell - 1\}} \mathbb{1}_{\{\gamma_{j'}^{(N)} \in (\frac{k}{n}, \frac{k+1}{n}]\}} \mathbb{1}_{\{\kappa_{j'}^{(N)} = \kappa_j^{(N)}\}}.
\end{aligned} \tag{4.39}$$

For $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, using computations analogous to the derivation of the limit of $\mathcal{I}_1^{(n)}$ in (4.27), the limit of the first term on the right-hand side of (4.39) is

$$\begin{aligned}
\lim_{n\to\infty} & \sum_{k=0}^{\lfloor nt \rfloor} \Bigl[\Bigl\langle \varphi\Bigl(\cdot + \frac{1}{n}, \frac{k}{n}\Bigr) - \varphi\Bigl(\cdot, \frac{k}{n}\Bigr), \nu_\ell^{(N)}\Bigl(\frac{k}{n}\Bigr)\Bigr\rangle\Bigr] \\
& = \int_0^t \langle \varphi_x(\cdot, s), \nu_\ell^{(N)}(s)\rangle\, ds.
\end{aligned} \tag{4.40}$$

For the second term, setting $\beta_{j,n}^{(N)} := \frac{1}{n}\lfloor n\beta_j^{(N)} \rfloor$, noting that $\beta_{j,n}^{(N)} \uparrow \beta_j^{(N)}$ as $n \to \infty$, using the continuity of φ and $a_j^{(N)}$, and the identity $a_j^{(N)}(\beta_j^{(N)}) = v_j$, we have

$$
\begin{aligned}
&\lim_{n\to\infty}\sum_{k=0}^{\lfloor nt\rfloor}\sum_{j}\varphi\Big(a_j^{(N)}\Big(\frac{k}{n}\Big)+\frac{1}{n},\frac{k}{n}\Big)\mathbb{1}_{\{\beta_j^{(N)}\in(\frac{k}{n},\frac{k+1}{n}]\}}\mathbb{1}_{\{\chi_j^{D,(N)}\geq\ell\}}\\
&\quad=\lim_{n\to\infty}\sum_{j}\sum_{k=0}^{\lfloor nt\rfloor}\varphi\Big(a_j^{(N)}\Big(\frac{k}{n}\Big)+\frac{1}{n},\frac{k}{n}\Big)\mathbb{1}_{\{\beta_j^{(N)}\in(\frac{k}{n},\frac{k+1}{n}]\}}\mathbb{1}_{\{\chi_j^{D,(N)}\geq\ell\}}\\
&\quad=\lim_{n\to\infty}\sum_{j}\varphi\Big(a_j^{(N)}(\beta_{j,n}^{(N)})+\frac{1}{n},\beta_{j,n}^{(N)}\Big)\mathbb{1}_{\{\beta_j^{(N)}\leq\frac{\lceil nt\rceil}{n}\}}\mathbb{1}_{\{\chi_j^{D,(N)}\geq\ell\}}\\
&\quad=\sum_{j}\varphi(v_j,\beta_j^{(N)})\mathbb{1}_{\{\beta_j^{(N)}\leq t\}}\mathbb{1}_{\{\chi_j^{D,(N)}\geq\ell\}}\\
&\quad=\mathcal{D}_{\varphi,\ell}^{(N)}(t), \qquad (4.41)
\end{aligned}
$$

where the last equality follows from (4.18).

Likewise, for the third term, setting $\alpha_{j,n}^{(N)}:=\frac{1}{n}\lfloor n\alpha_j^{(N)}\rfloor$, and noting that $a_j^{(N)}(\alpha_j^{(N)})=0$ and $\alpha_{j,n}^{(N)}\uparrow\alpha_j^{(N)}$ as $n\to\infty$, we obtain

$$
\begin{aligned}
&\lim_{n\to\infty}\sum_{k=0}^{\lfloor nt\rfloor}\sum_{j}\varphi\Big(a_j^{(N)}\Big(\frac{k+1}{n}\Big),\frac{k}{n}\Big)\mathbb{1}_{\{\alpha_j^{(N)}\in(\frac{k}{n},\frac{k+1}{n}]\}}\mathbb{1}_{\{\chi_j^{K,(N)}\geq\ell\}}\\
&\quad=\lim_{n\to\infty}\sum_{j}\varphi\Big(a_j^{(N)}\Big(\alpha_{j,n}^{(N)}+\frac{1}{n}\Big),\alpha_{j,n}^{(N)}\Big)\mathbb{1}_{\{0\leq\alpha_j^{(N)}\leq\frac{\lceil nt\rceil}{n}\}}\mathbb{1}_{\{\chi_j^{K,(N)}\geq\ell\}}\\
&\quad=\sum_{j}\varphi(0,\alpha_j^{(N)})\mathbb{1}_{\{0\leq\alpha_j^{(N)}\leq t\}}\mathbb{1}_{\{\chi_j^{K,(N)}\geq\ell\}}. \qquad (4.42)
\end{aligned}
$$

Further simplification of (4.42) is slightly different for $\ell=1$ and $\ell\geq 2$. When $\ell\geq 2$, due to the nonidling assumption, the service entry time of any job to a queue of length at least ℓ just after service entry coincides with the departure time of another job from the same queue, which had length at least $\ell+1$ just before the departure. Therefore, for $\ell\geq 2$, by definition (4.16) of $D_\ell^{(N)}$ we have

$$
\begin{aligned}
\sum_{j}\varphi(0,\alpha_j^{(N)})\mathbb{1}_{\{0\leq\alpha_j^{(N)}\leq t\}}\mathbb{1}_{\{\chi_j^{K,(N)}\geq\ell\}}&=\sum_{j}\varphi(0,\beta_j^{(N)})\mathbb{1}_{\{\beta_j^{(N)}\leq t\}}\mathbb{1}_{\{\chi_j^{D,(N)}\geq\ell+1\}}\\
&=\int_{[0,t]}\varphi(0,s)\,dD_{\ell+1}^{(N)}(s). \qquad (4.43)
\end{aligned}
$$

To simplify (4.42) for $\ell=1$, we first define the cumulative service entry process:

$$
K^{(N)}(t):=\sum_{j=j_0}^{\infty}\mathbb{1}_{\{0\leq\alpha_j^{(N)}\leq t\}},\qquad t\geq 0. \qquad (4.44)
$$

LEMMA 4.8. *Given $D_2^{(N)}$ and $E^{(N),i}$ defined in* (4.16) *and* (4.10), *respectively*,

$$K^{(N)}(t) = D_2^{(N)}(t) + \sum_{i=1}^{N} \int_0^t \mathbb{1}_{\{X^{(N),i}(u-)=0\}} \, dE^{(N),i}(u), \qquad t \geq 0. \tag{4.45}$$

PROOF. Service entries can be classified into two types, based on whether or not the queue was empty right before service entry. Thus, we can expand (4.44) as

$$K^{(N)}(t) = \sum_j \mathbb{1}_{\{0 \leq \alpha_j^{(N)} \leq t\}} \mathbb{1}_{\{\chi_j^{(N)}(\alpha_j^{(N)}-) \geq 1\}} + \sum_j \mathbb{1}_{\{0 \leq \alpha_j^{(N)} \leq t\}} \mathbb{1}_{\{\chi_j^{(N)}(\alpha_j^{(N)}-) = 0\}}.$$

Due to the nonidling assumption, the service entry time of the first type coincides with the departure time of another job from the same queue, which had length of at least 2 just before departure. On the other hand, a service entry time of the second type coincides with the arrival time of the same job to an empty queue. Recalling that $\kappa_j^{(N)}$ is the queue index of job j, we can then write

$$\begin{aligned} K^{(N)}(t) &= \sum_j \mathbb{1}_{\{\beta_j^{(N)} \leq t\}} \mathbb{1}_{\{\chi_j^{D,(N)} \geq 2\}} + \sum_{j \geq 1} \mathbb{1}_{\{\gamma_j^{(N)} \leq t\}} \mathbb{1}_{\{\chi_j^{E,(N)} = 1\}} \\ &= \sum_j \mathbb{1}_{\{\beta_j^{(N)} \leq t\}} \mathbb{1}_{\{\chi_j^{D,(N)} \geq 2\}} \\ &\quad + \sum_{i=1}^{N} \sum_{j=1}^{\infty} \mathbb{1}_{\{X^{(N),i}(\gamma_j^{(N)}-)=0\}} \mathbb{1}_{\{\gamma_j^{(N)} \leq t\}} \mathbb{1}_{\{\kappa_j^{(N)} = i\}}. \end{aligned}$$

Equation (4.45) then follows from (4.16) and (4.10). □

By (4.44) and (4.45), the fact that $\mathbb{1}_{\{\chi_j^{K,(N)} \geq 1\}} = 1$ (see Remark 4.2), and (4.14), we have

$$\begin{aligned} &\sum_j \varphi(0, \alpha_j^{(N)}) \mathbb{1}_{\{0 \leq \alpha_j^{(N)} \leq t\}} \mathbb{1}_{\{\chi_j^{K,(N)} \geq 1\}} \\ &\quad = \int_{[0,t]} \varphi(0,u) \, dD_2^{(N)}(u) + \sum_{i=1}^{N} \int_0^t \varphi(0,u) \mathbb{1}_{\{X^{(N),i}(u-)=0\}} \, dE^{(N),i}(u) \\ &\quad = \int_{[0,t]} \varphi(0,u) \, dD_2^{(N)}(u) + \mathcal{R}_{\varphi,1}^{(N)}(t). \end{aligned} \tag{4.46}$$

Finally, the last term on the right-hand side of (4.39) is zero for $\ell = 1$. For $\ell \geq 2$, changing the order of summation, setting $\gamma_{j',n}^{(N)} = \frac{1}{n} \lfloor n \gamma_{j'}^{(N)} \rfloor$, noting that on Ω_t, the arrival of j' is the only event taking place in the interval $(k/n, (k+1)/n]$, the limit of the last term on the right-hand side of (4.39) is equal to

$$\lim_{n\to\infty} \sum_{j'\geq 1} \sum_{j} \varphi\Big(a_j^{(N)}\Big(\gamma_{j',n}^{(N)} + \frac{1}{n}\Big), \gamma_{j',n}^{(N)}\Big) \mathbb{1}_{\{\alpha_j^{(N)} \leq \gamma_{j',n}^{(N)}\}} \mathbb{1}_{\{\beta_j^{(N)} > \gamma_{j',n}^{(N)}\}}$$
$$\times \mathbb{1}_{\{\chi_j^{(N)}(\gamma_{j',n}^{(N)})=\ell-1\}} \mathbb{1}_{\{\gamma_{j'}^{(N)} \leq \frac{\lceil nt\rceil}{n}\}} \mathbb{1}_{\{\kappa_{j'}^{(N)}=\kappa_j^{(N)}\}}.$$

Since $\gamma_{j',n}^{(N)} \uparrow \gamma_{j'}^{(N)}$ as $n \to \infty$, the fact that on Ω_t, $\alpha_j^{(N)}, \beta_j^{(N)} \neq \gamma_{j'}^{(N)}$ for all $j \neq j'$, and by the continuity of φ and $a_j^{(N)}$, the last display is equal to

$$\sum_{j'\geq 1} \sum_{j} \varphi(a_j^{(N)}(\gamma_{j'}^{(N)}), \gamma_{j'}^{(N)}) \mathbb{1}_{\{\alpha_j^{(N)} < \gamma_{j'}^{(N)} \leq t \wedge \beta_j^{(N)}\}} \mathbb{1}_{\{\chi_j^{(N)}(\gamma_{j'}^{(N)}-)=\ell-1\}} \mathbb{1}_{\{\kappa_{j'}^{(N)}=\kappa_j^{(N)}\}}.$$

Partitioning jobs in terms of their queues, and using (4.11), the last display equals

$$\sum_{i=1}^{N} \sum_{j'\geq 1} \mathbb{1}_{\{\kappa_{j'}^{(N)}=i\}} \mathbb{1}_{\{\gamma_{j'}^{(N)} \leq t\}} \sum_{j} \varphi(a_j^{(N)}(\gamma_{j'}^{(N)}), \gamma_{j'}^{(N)}) \mathbb{1}_{\{\alpha_j^{(N)} < \gamma_{j'}^{(N)}\}}$$
$$\times \mathbb{1}_{\{\beta_j^{(N)} \geq \gamma_{j'}^{(N)}\}} \mathbb{1}_{\{\chi_j^{(N)}(\gamma_{j'}^{(N)}-)=\ell-1\}} \mathbb{1}_{\{\kappa_j^{(N)}=i\}}$$
$$= \sum_{i=1}^{N} \sum_{j'\geq 1} \mathbb{1}_{\{\kappa_{j'}^{(N)}=i\}} \mathbb{1}_{\{\gamma_{j'}^{(N)} \leq t\}} \langle \varphi(\cdot, \gamma_{j'}^{(N)}), \nu_{\ell-1}^{(N),i}(\gamma_{j'}^{(N)}-)$$
$$- \nu_{\ell}^{(N),i}(\gamma_{j'}^{(N)}-)\rangle$$
$$= \sum_{i=1}^{N} \int_{(0,t]} \langle \varphi(\cdot, s), \nu_{\ell-1}^{(N),i}(s-) - \nu_{\ell}^{(N),i}(s-)\rangle \, dE^{(N),i}(s)$$
$$= \mathcal{R}_{\varphi,\ell}^{(N)}(t), \tag{4.47}$$

where $E^{(N),i}$ and $\mathcal{R}_{\varphi,\ell}^{(N)}(t)$ are defined in (4.10) and (4.13), respectively.

We now combine the above observations to conclude the proof.

PROOF OF PROPOSITION 4.5. Equation (4.24) follows from (4.26), (4.29), (4.39)–(4.43), (4.46) and (4.47). To establish (4.25), note that for $\varphi = \mathbf{1}$, $\mathcal{I}_1^{(n)}(t)$ and the first term on the right-hand side of (4.39) are zero for all $t \geq 0$. Since Lemma 4.6 and the calculation of other terms on the right-hand side of (4.39) are valid for all $\varphi \in \mathbb{C}_b([0, L) \times [0, \infty))$, (4.25) follows on setting $\varphi = \mathbf{1}$ in (4.24). □

REMARK 4.9. Equation (4.24) remains valid for functions φ on $[0, L) \times \mathbb{R}_+$ of the form $\varphi(x, s) = f(x)$ for some $f \in \mathbb{C}_c^1[0, L)$, even though they are not compactly supported on $[0, L) \times \mathbb{R}_+$ (because the compact support condition on the s variable is only used in the computation of $\mathcal{I}_1$ in (4.29), which is zero for functions of the above form). This property is used in the proof of Proposition 6.15.

5. Martingale decomposition for routing and departure processes. Fix $N \in \mathbb{N}$. In Section 5.1, we state a martingale decomposition result for the φ-weighted routing process $\mathcal{R}^{(N)}_{\varphi,\ell}$ and departure process $\mathcal{D}^{(N)}_{\varphi,\ell}$ defined in (4.13) and (4.18), respectively. The proofs are given in Section 5.3, and rely on an alternative characterization of the filtration $\{\mathcal{F}^{(N)}_t\}$ in terms of a marked point process introduced in Section 5.2. Unlike Proposition 4.5, these results are specific to our assumptions on the arrival process and load balancing algorithm, although the general method can be adapted to analyze other models.

5.1. *The form of compensators.* For $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, define

$$B^{(N)}_{\varphi,1}(t) := \int_0^t h^{(N)}_E\big(R^{(N)}_E(u)\big)\varphi(0,u)\big(1 - \big(\overline{S}^{(N)}_1(u)\big)^d\big)\,du, \tag{5.1}$$

where $h^{(N)}_E$ is the hazard rate of the interarrival distribution, and for $\ell \geq 2$, set

$$\begin{aligned} B^{(N)}_{\varphi,\ell}(t) := \int_0^t & h^{(N)}_E\big(R^{(N)}_E(u)\big)\mathfrak{P}_d\big(\overline{S}^{(N)}_{\ell-1}(u), \overline{S}^{(N)}_\ell(u)\big) \\ & \times \big\langle \varphi(\cdot,u), \overline{\nu}^{(N)}_{\ell-1}(u) - \overline{\nu}^{(N)}_\ell(u)\big\rangle\,du. \end{aligned} \tag{5.2}$$

PROPOSITION 5.1. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, for* $\ell \geq 1$ *and* $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, *the process*

$$\mathcal{N}^{(N)}_{\varphi,\ell} := \mathcal{R}^{(N)}_{\varphi,\ell} - B^{(N)}_{\varphi,\ell}, \tag{5.3}$$

is a local $\{\mathcal{F}^{(N)}_t\}$-*martingale, with quadratic variation*

$$\big[\mathcal{N}^{(N)}_{\varphi,\ell}\big](t) = \mathcal{R}^{(N)}_{\varphi^2,\ell}(t), \qquad t \geq 0. \tag{5.4}$$

Proposition 5.1 is a key result and its proof is given in Section 5.3. Next, for $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, define

$$A^{(N)}_{\varphi,\ell}(t) := \int_0^t \big\langle \varphi(\cdot,s)h(\cdot), \nu^{(N)}_\ell(s)\big\rangle\,ds, \qquad \forall t \geq 0. \tag{5.5}$$

PROPOSITION 5.2. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, for* $\ell \geq 1$ *and* $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, *the process*

$$\mathcal{M}^{(N)}_{\varphi,\ell} := \mathcal{D}^{(N)}_{\varphi,\ell} - A^{(N)}_{\varphi,\ell}, \tag{5.6}$$

is a local $\{\mathcal{F}^{(N)}_t\}$-*martingale, with quadratic variation*

$$\big[\mathcal{M}^{(N)}_{\varphi,\ell}\big](t) = \mathcal{D}^{(N)}_{\varphi^2,\ell}(t), \qquad t \geq 0. \tag{5.7}$$

Since the proof of Proposition 5.2 is similar to (in fact much simpler than) that of Proposition 5.1, it is omitted (full details can be found in Appendix B of [5]). A similar result for a different model and filtration can also be found in [24], (5.24), (5.25) and Lemma 5.4.

REMARK 5.3. Substituting (5.3), (5.6) and (5.5) into (4.24), we have

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_\ell^{(N)}(t)\rangle = {} & \langle \varphi(\cdot,0), \nu_\ell^{(N)}(0)\rangle \\ & + \int_0^t \langle \varphi_s(\cdot,s) + \varphi_x(\cdot,s) - \varphi(\cdot,s)h(\cdot), \nu_\ell^{(N)}(s)\rangle\, ds \\ & + \int_{[0,t]} \varphi(0,s)\, dD_{\ell+1}^{(N)}(s) + B_{\varphi,\ell}^{(N)}(t) - \mathcal{M}_{\varphi,\ell}^{(N)}(t) + \mathcal{N}_{\varphi,\ell}^{(N)}(t).\end{aligned} \tag{5.8}$$

We now state an elementary lemma used in the proof of Proposition 5.1. Suppose $(\Omega, \mathcal{G}, \{\mathcal{G}_t\}, \mathbb{P})$ is a filtered probability space that satisfies the usual conditions, and let $\xi = \{\xi(t); t \geq 0\}$ be a point process adapted to $\{\mathcal{G}_t\}$. Recall that a nonnegative $\{\mathcal{G}_t\}$-progressive process $\{\varrho(t); t \geq 0\}$ is called a $\{\mathcal{G}_t\}$-intensity of ξ if for all $t \geq 0$, $\int_0^t \varrho(s)\, ds < \infty$ almost surely, and for every nonnegative $\{\mathcal{G}_t\}$-predictable processes H, $\mathbb{E}[\int_0^\infty H(t)\, d\xi(t)] = \mathbb{E}[\int_0^\infty H(s)\varrho(s)\, ds]$. The next result is elementary (e.g., it follows from Lemma II.L3 of [12] and equation (18.1), Chapter IV of [32], as elaborated in the Proof of Lemma 5.4 in [5]).

LEMMA 5.4. *Let ϱ be a $\{\mathcal{G}_t\}$-intensity of a point process ξ on $(\Omega, \mathcal{G}, \{\mathcal{G}_t\}, \mathbb{P})$, and given a locally bounded, $\{\mathcal{G}_t\}$-predictable process θ, define $\zeta(t) := \int_0^t \theta(s)\, d\xi(s)$. Then $\zeta(t) - \int_0^t \theta(s)\varrho(s)\, ds$, $t \geq 0$, is a local $\{\mathcal{G}_t\}$-martingale, with quadratic variation*

$$[\zeta](t) = \int_0^t \theta^2(s)\, d\xi(s), \qquad t \geq 0. \tag{5.9}$$

5.2. *A marked point process representation.* In this section, we construct a point process $\mathcal{T}^{(N)}$ consisting of all arrival and departure times, marked by their type and their corresponding queue index. This point process has the property that its natural filtration, together with the σ-algebra generated by initial conditions, is equivalent to the filtration $\{\mathcal{F}_t^{(N)}; t \geq 0\}$ defined in (4.22). Moreover, each auxiliary process defined in Section 4.2 can be represented as an integral with respect to $\mathcal{T}^{(N)}$, which allows us to more easily identify its compensator.

Consider the set

$$\mathbb{T}^{(N)} := \{(\gamma_j^{(N)}, (\mathfrak{E}, \kappa_j^{(N)})); j \geq 1\} \cup \{(\beta_j^{(N)}, (\mathfrak{D}, \kappa_j^{(N)})); j \geq j_0\},$$

which is the union of all arrival times $\gamma_j^{(N)}$, marked by the tag $\mathfrak{E}$ (indicating that it is an arrival time) and the index of the queue to which job j is routed, and all departure times $\beta_j^{(N)}$, marked by the tag $\mathfrak{D}$ (indicating that it is a departure time) and the index of the queue from which job j departed. Since the interarrival and service distributions $G_E^{(N)}$ and G are absolutely continuous with respect to Lebesgue measure, by Assumptions I and II(a), almost surely at most one arrival to and at most one departure from each queue can occur at any given time. Let $\tau_0 := 0$ and z_0 be a constant (whose value is irrelevant), and define the sequence of *events* $\{(\tau_k^{(N)}, z_k^{(N)}); k \geq 1\}$, each composed of an *event time* $\tau_k^{(N)}$ and an *event mark* $z_k^{(N)}$, to be the relabeling (i.e., a one-to-one correspondence) of $\mathbb{T}^N$ sorted by lexicographic order, assuming $\mathfrak{D} < \mathfrak{E}$. That is, events are ordered first by event times ($\tau_k^{(N)} \leq \tau_{k+1}^{(N)}$), then by event type (departure first, then arrival) and finally by queue index (with smaller indices first). Let $\mathcal{T}^{(N)} = \{\mathcal{T}^{(N)}(t); t \geq 0\}$ be the corresponding marked point process. Clearly, for every index $i \in \{1, \dots, N\}$,

$$\mathcal{T}^{(N)}(\mathfrak{E}, i; t) := \sum_{k\geq 1} \mathbb{1}_{\{\tau_k^{(N)} \leq t\}} \mathbb{1}_{\{z_k^{(N)} = (\mathfrak{E}, i)\}} = E^{(N),i}(t) \tag{5.10}$$

and

$$\mathcal{T}^{(N)}(\mathfrak{D}, i; t) := \sum_{k\geq 1} \mathbb{1}_{\{\tau_k^{(N)} \leq t\}} \mathbb{1}_{\{z_k^{(N)} = (\mathfrak{D}, i)\}} = D^{(N),i}(t). \tag{5.11}$$

These relations show that the filtration $\{\mathcal{F}_t^{(N)}\}$ in (4.22) has the representation

$$\mathcal{F}_t^{(N)} = \sigma(I_0) \vee \mathcal{F}_t^{\mathcal{T},(N)}, \qquad t \geq 0, \tag{5.12}$$

where $\{\mathcal{F}_t^{\mathcal{T},(N)}\}$ is the filtration generated by the marked point process $\mathcal{T}^{(N)}$; see equation (1.2), page 57 of [12].

At any time t, server i is called *busy* if $X^{(N),i}(t) \geq 1$, and is called *idle* otherwise. By the nonidling assumption, there is a job receiving service at any busy server i at time t, and we define $a^{(N),i}(t)$ to be the age of that job. Using this notation, for $\ell \geq 1$, we can rewrite the definition (4.11) of $\nu_\ell^{(N),i}$ as

$$\nu_\ell^{(N),i}(t) = \mathbb{1}_{\{X^{(N),i}(t) \geq \ell\}} \delta_{a^{(N),i}(t)}, \tag{5.13}$$

which when combined with (4.12), yields

$$\nu_\ell^{(N)}(t) = \sum_{i=1}^{N} \mathbb{1}_{\{X^{(N),i}(t) \geq \ell\}} \delta_{a^{(N),i}(t)}. \tag{5.14}$$

For $k \geq 0$, define

$$\mathfrak{B}_k^{(N)} := \{i : X^{(N),i}(\tau_k^{(N)}) \geq 1\} \tag{5.15}$$

to be the set of busy servers at time $\tau_k^{(N)}$, and note that for $i \in \mathfrak{B}_k^{(N)}$, $a^{(N),i}(\tau_k^{(N)})$ is well defined. Define $\xi_{k+1}^{(N)}$ to be the next arrival time strictly after $\tau_k^{(N)}$, and for $i = 1, \dots, N$, define $\sigma_{k+1}^{(N),i}$ to be the next time strictly after $\tau_k^{(N)}$ when there is a departure from queue i if $i \in \mathfrak{B}_k^{(N)}$, and $\sigma_{k+1}^{(N),i} = \infty$, otherwise. When the event time $\tau_k^{(N)}$ is distinct, that is, $\tau_k^{(N)} \neq \tau_{k'}^{(N)}$ for all $k' \neq k$, the next event time will be the minimum among the first arrival time after τ_k and the next departure time from queues that are busy at time τ_k. Therefore, defining

$$\tilde{\Omega}_k := \{\omega \in \Omega; \tau_k \neq \tau_{k'}, \forall k' \neq k\}, \tag{5.16}$$

to be the set of realizations on which the event time $\tau_k^{(N)}$ is distinct, we have

$$\tau_{k+1}^{(N)} = \min(\xi_{k+1}^{(N)}, \sigma_{k+1}^{(N),i}; i = 1, \dots, N) \qquad \text{on } \tilde{\Omega}_k. \tag{5.17}$$

The next lemma identifies the joint distribution of the next arrival and departure times given $\mathcal{F}_{\tau_k^{(N)}}^{(N)}$.

LEMMA 5.5. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, for* $k \geq 0$, $\mathbb{P}\{\tilde{\Omega}_k\} = 1$, $\xi_{k+1}^{(N)}$ *and* $\sigma_{k+1}^{(N),i}$, $i = 1, \dots, N$, *are conditionally independent given* $\mathcal{F}_{\tau_k^{(N)}}^{(N)}$, *and*

$$\mathbb{P}\{\xi_{k+1}^{(N)} - \tau_k^{(N)} > b | \mathcal{F}_{\tau_k^{(N)}}^{(N)}\} = \frac{\overline{G}_E^{(N)}(R_E^{(N)}(\tau_k^{(N)}) + b)}{\overline{G}_E^{(N)}(R_E^{(N)}(\tau_k^{(N)}))}, \qquad b > 0, \tag{5.18}$$

and for $i = 1, \dots, N$,

$$\begin{aligned} &\mathbb{1}_{\{i \in \mathfrak{B}_k^{(N)}\}} \mathbb{P}\{\sigma_{k+1}^{(N),i} - \tau_k^{(N)} > b | \mathcal{F}_{\tau_k^{(N)}}^{(N)}\} \\ &\quad = \mathbb{1}_{\{i \in \mathfrak{B}_k^{(N)}\}} \frac{\overline{G}(a^{(N),i}(\tau_k^{(N)}) + b)}{\overline{G}(a^{(N),i}(\tau_k^{(N)}))}, \qquad b > 0. \end{aligned} \tag{5.19}$$

The result in Lemma 5.5 is intuitive and follows from the independence of the interarrival and service times. However, a completely rigorous proof is rather involved and technical, although involving fairly routine calculations. Hence, we omit the proof, and refer the reader to Appendix A of [5] for all details. Using Lemma 5.5, we can rewrite (5.17) as

$$\tau_{k+1}^{(N)} = \min(\xi_{k+1}^{(N)}, \sigma_{k+1}^{(N),i}; i = 1, \dots, N), \qquad \text{a.s.} \tag{5.20}$$

We now state a consequence of Lemma 5.5. Recall that a sequence $\{t_n; n \in \mathbb{N}\}$ is called nonexplosive if for every $T < \infty$, there are finitely many n with $t_n \leq T$.

COROLLARY 5.6. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, almost surely, the sequence of event times* $\{\tau_k^{(N)}; k \geq 0\}$ *is nonexplosive, and* $\mathbb{P}\{\Omega_t\} = 1$, $t \geq 0$.

PROOF. Fix $t \geq 0$, define $\hat{\Omega}_t = \hat{\Omega}_t^{(N)} = \{\omega : X^{(N)}(0) < \infty, E^{(N)}(t) < \infty\}$. Since $X^{(N)}(0) < \infty$ almost surely by Assumption III(a) and $E^{(N)}$ is a renewal process with nondegenerate interarrival time distribution by Assumption I, $\mathbb{P}\{\hat{\Omega}_t\} = 1$. Moreover, $D^{(N)}(t)$ is also finite on $\hat{\Omega}_t$ by (4.17). Thus, on $\hat{\Omega}_t$, and hence, almost surely, the total number of events up to t is finite and $\{\tau_k^{(N)}\}$ is nonexplosive.

Moreover, the set $\tilde{\Omega} := \bigcup_{k \geq 0} \tilde{\Omega}_k$ of realizations on which all events are distinct has full measure by Lemma 5.5. For every $\omega \in \hat{\Omega}_t \cap \tilde{\Omega}$, the quantity $\Delta(\omega) = \Delta^{(N)}(\omega) := \inf_{k:\tau_k^{(N)} \leq t}(\tau_{k+1}^{(N)} - \tau_k^{(N)})$, is strictly positive because it is the infimum of finitely many positive numbers. This means that for $n > 1/\Delta$, the distance between any two events prior to time t exceeds $1/n$. Therefore, $\omega \in \Omega_{\frac{k}{n},\frac{1}{n}}$ for all $k = 0, \ldots, \lfloor nt \rfloor$, and hence, $\omega \in \Omega_t$. This implies $\hat{\Omega}_t \cap \tilde{\Omega} \subseteq \Omega_t$, and hence, $\mathbb{P}\{\Omega_t\} = 1$. □

5.3. *Compensator for the weighted routing measure.* We state our first result.

LEMMA 5.7. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, for* $i = 1, \ldots, N$, *the process* $E^{(N),i}$ *defined in* (4.10) *has the following* $\{\mathcal{F}_t^{(N)}\}$-*intensity process*:

$$\left\{\frac{1}{N} h_E^{(N)}\big(R_E^{(N)}(t-)\big) \sum_{\ell=1}^{\infty} \mathbb{1}_{\{X^{(N),i}(t-)=\ell-1\}} \mathfrak{P}_d\big(\overline{S}_{\ell-1}^{(N)}(t-), \overline{S}_\ell^{(N)}(t-)\big); t \geq 0\right\}. \tag{5.21}$$

PROOF. Throughout this proof, we omit the superscript (N) for ease of notation. Suppose that for $k \geq 0$ and $i = 1, \ldots, N$, the conditional density $f_{k+1}^{\mathfrak{E},i}$ defined by

$$\mathbb{P}\{\tau_{k+1} - \tau_k \in A, z_{k+1} = (\mathfrak{E}, i) | \mathcal{F}_{\tau_k}\} = \int_A f_{k+1}^{\mathfrak{E},i}(\omega, r)\, dr, \\ \omega \in \Omega, A \in \mathcal{B}[0, \infty), \tag{5.22}$$

exists. Then, by the representation in (5.10) of E^i in terms of $\mathcal{T}^{(N)}$, and the fact that the sequence $\{\tau_k\}$ is nonexplosive by Corollary 5.6, it follows from Theorem III.T7 of [12], comment (β) and (2.10), that the process

$$\sum_{k=0}^{\infty} \frac{f_{k+1}^{\mathfrak{E},i}(\omega, t - \tau_k)}{\mathbb{P}\{\tau_{k+1} > t | \mathcal{F}_{\tau_k}\}} \mathbb{1}_{\{\tau_k < t \leq \tau_{k+1}\}} \tag{5.23}$$

is an $\{\mathcal{F}_t\}$-intensity of E^i.

We now show that $f_{k+1}^{\mathfrak{E},i}$ in (5.22) exists. First, note that by (5.20), the next event after τ_k is an arrival to queue i, that is, $z_{k+1} = (\mathfrak{E}, i)$, if the next arrival occurs before the next departure from any queue and the arriving job, which has index

$E(\tau_{k+1})$, is routed to queue i. Hence, defining $\sigma_{k+1} := \min(\sigma_{k+1}^{i'}; i' = 1, \dots, N)$, we have

$$\begin{aligned}
&\mathbb{P}\{\tau_{k+1} - \tau_k > t, z_{k+1} = (\mathfrak{E}, i) | \mathcal{F}_{\tau_k}\} \\
&\quad = \mathbb{P}\{\xi_{k+1} > \tau_k + t, \sigma_{k+1} > \xi_{k+1}, \kappa_{E(\tau_{k+1})} = i | \mathcal{F}_{\tau_k}\} \\
&\quad = \mathbb{E}\big[\mathbb{1}_{\{\xi_{k+1} > \tau_k + t, \sigma_{k+1} > \xi_{k+1}\}} \mathbb{P}\{\kappa_{E(\tau_{k+1})} = i | \mathcal{F}_{\tau_k}, \xi_{k+1}, \sigma_{k+1}\} | \mathcal{F}_{\tau_k}\big]. \qquad (5.24)
\end{aligned}$$

The queue to which job $j = E(\tau_{k+1})$ is routed is given by (4.2), which is a function of the random queue choices vector ι_j, queue lengths at time τ_k and random tie-breakers. According to (4.2), this job is routed to a queue of length exactly $\ell - 1$, if and only if all selected queue indices ι_j have lengths at least $\ell - 1$, and at least one of them has length exactly $\ell - 1$. Since ι_j is independent of all other random variables, the conditional probability (given $\mathcal{F}_{\tau_k}$, σ_{k+1} and ξ_{k+1}) that the job is routed to a queue of length $\ell - 1$ is $(\overline{S}_{\ell-1}(\tau_k))^d - (\overline{S}_\ell(\tau_k))^d$. Moreover, the job is equally likely to be routed to any queue of length $\ell - 1$. Since there are $S_{\ell-1}(\tau_k) - S_\ell(\tau_k)$ such queues, on the event $X^i(\tau_k) = \ell - 1$, we have

$$\begin{aligned}
\mathbb{P}\{\kappa_{E(\tau_{k+1})} = i | \mathcal{F}_{\tau_k}, \xi_{k+1}, \sigma_{k+1}\} &= \frac{(\overline{S}_{\ell-1}(\tau_k))^d - (\overline{S}_\ell(\tau_k))^d}{S_{\ell-1}(\tau_k) - S_\ell(\tau_k)} \\
&= \frac{1}{N} \mathfrak{P}_d\big(\overline{S}_{\ell-1}(\tau_k), \overline{S}_\ell(\tau_k)\big),
\end{aligned}$$

where the polynomial $\mathfrak{P}_d$ is defined in (2.8). Therefore,

$$\begin{aligned}
&\mathbb{P}\{\kappa_{E(\tau_{k+1})} = i | \mathcal{F}_{\tau_k}, \xi_{k+1}, \sigma_{k+1}\} \\
&\quad = \sum_{\ell=1}^{\infty} \mathbb{1}_{\{X^i(\tau_k) = \ell - 1\}} \mathbb{P}\{\kappa_{E(\tau_{k+1})} = i | \mathcal{F}_{\tau_k}, \xi_{k+1}, \sigma_{k+1}\} \\
&\quad = \frac{1}{N} \sum_{\ell=1}^{\infty} \mathbb{1}_{\{X^i(\tau_k) = \ell - 1\}} \mathfrak{P}_d\big(\overline{S}_{\ell-1}(\tau_k), \overline{S}_\ell(\tau_k)\big). \qquad (5.25)
\end{aligned}$$

Moreover, using (5.18) and Lemma 5.5, we have

$$\begin{aligned}
&\mathbb{P}\{\xi_{k+1} > \tau_k + t, \sigma_{k+1} > \xi_{k+1} | \mathcal{F}_{\tau_k}\} \\
&\quad = \frac{1}{\overline{G}_E(R_E(\tau_k))} \int_t^{\infty} \mathbb{P}\{\sigma_{k+1} - \tau_k > s | \mathcal{F}_{\tau_k}\} g_E\big(s + R_E(\tau_k)\big)\, ds. \qquad (5.26)
\end{aligned}$$

Therefore, by (5.24)–(5.26) and the fact that the right-hand side of (5.25) is $\mathcal{F}_{\tau_k}$-measurable, for $k \in \mathbb{Z}_+$, $i \in \{1, \dots, N\}$, the density $f_{k+1}^{\mathfrak{E},i}$ exists, and for $t \geq \tau_k$,

$$\begin{aligned}
f_{k+1}^{\mathfrak{E},i}(t - \tau_k) &= \mathbb{P}\{\sigma_{k+1} > t | \mathcal{F}_{\tau_k}\} \frac{g_E(t - \tau_k + R_E(\tau_k))}{N \overline{G}_E(R_E(\tau_k))} \\
&\quad \times \sum_{\ell=1}^{\infty} \mathbb{1}_{\{X^i(\tau_k) = \ell - 1\}} \mathfrak{P}_d\big(\overline{S}_{\ell-1}(\tau_k), \overline{S}_\ell(\tau_k)\big). \qquad (5.27)
\end{aligned}$$

Similarly, by (5.20), Lemma 5.5 and (5.18), for every $k \geq 0$ and $t \geq \tau_k$ we have

$$\begin{aligned}\mathbb{P}\{\tau_{k+1} > t|\mathcal{F}_{\tau_k}\} &= \mathbb{P}\{\xi_{k+1} \wedge \sigma_{k+1} > t|\mathcal{F}_{\tau_k}\} \\ &= \mathbb{P}\{\sigma_{k+1} > t|\mathcal{F}_{\tau_k}\}\mathbb{P}\{\xi_{k+1} > t|\mathcal{F}_{\tau_k}\} \\ &= \mathbb{P}\{\sigma_{k+1} > t|\mathcal{F}_{\tau_k}\}\frac{\overline{G}_E(t-\tau_k+R_E(\tau_k))}{\overline{G}_E(R_E(\tau_k))}. \end{aligned} \tag{5.28}$$

Combining (5.27) and (5.28) with (5.23), recalling that $h_E = g_E/\overline{G}_E$ is the hazard rate function of the interarrival times, and noting that on $(\tau_k, \tau_{k+1}]$, X^i, S_ℓ are constant and R_E has a unit slope, we see that

$$\begin{aligned}&\frac{1}{N}\sum_{k=0}^{\infty} h_E\bigl(t-\tau_k+R_E(\tau_k)\bigr) \\ &\qquad \times \sum_{\ell=1}^{\infty} \mathbb{1}_{\{X^i(\tau_k)=\ell-1\}}\mathfrak{P}_d\bigl(\overline{S}_{\ell-1}(\tau_k), \overline{S}_\ell(\tau_k)\bigr)\mathbb{1}_{\{\tau_k<t\leq\tau_{k+1}\}} \\ &\quad = \frac{1}{N}\sum_{k=0}^{\infty} h_E\bigl(R_E(t-)\bigr) \\ &\qquad \times \sum_{\ell=1}^{\infty} \mathbb{1}_{\{X^i(t-)=\ell-1\}}\mathfrak{P}_d\bigl(\overline{S}_{\ell-1}(t-), \overline{S}_\ell(t-)\bigr)\mathbb{1}_{\{\tau_k<t\leq\tau_{k+1}\}} \\ &\quad = \frac{1}{N} h_E\bigl(R_E(t-)\bigr) \\ &\qquad \times \sum_{\ell=1}^{\infty} \mathbb{1}_{\{X^i(t-)=\ell-1\}}\mathfrak{P}_d\bigl(\overline{S}_{\ell-1}(t-), \overline{S}_\ell(t-)\bigr), \qquad t \geq 0,\end{aligned} \tag{5.29}$$

is a $\{\mathcal{F}_t\}$-intensity of $E^{(N),i}$. □

We now use Lemma 5.7 to prove the martingale decomposition for $\mathcal{R}^{(N)}_{\varphi,\ell}$.

PROOF OF PROPOSITION 5.1. We only give the proof for $\ell \geq 2$. The proof for $\ell = 1$ is similar; see the proof of Proposition 5.1 in [5] for details. Note from (4.13) that

$$\mathcal{R}^{(N)}_{\varphi,\ell} = \sum_{i=1}^{N} \mathcal{R}^{(N),i}_{\varphi,\ell}, \tag{5.30}$$

where

$$\mathcal{R}^{(N),i}_{\varphi,\ell}(t) := \begin{cases} \displaystyle\int_{(0,t]} \varphi(0,s)\mathbb{1}_{\{X^{(N),i}(s-)=0\}}\, dE^{(N),i}(s) \\ \qquad \text{if } \ell = 1, \\ \displaystyle\int_{(0,t]} \langle \varphi(\cdot,s), \nu^{(N),i}_{\ell-1}(s-) - \nu^{(N),i}_{\ell}(s-)\rangle\, dE^{(N),i}(s) \\ \qquad \text{if } \ell \geq 2. \end{cases} \tag{5.31}$$

Consider the setup of Lemma 5.4 with ξ, $\{\mathcal{G}_t\}$ as above, but with $\theta(s)$ replaced by

$$\langle \varphi(\cdot,s), \nu^{(N),i}_{\ell-1}(s-) - \nu^{(N),i}_{\ell}(s-)\rangle = \varphi(a^{(N),i}(s-),s)\mathbb{1}_{\{X^{(N),i}(s-)=\ell-1\}}$$

and hence, ζ replaced by $\mathcal{R}^{(N),i}_{\varphi,\ell}$. Since $a^{(N),i}$ and $X^{(N),i}$ are $\{\mathcal{F}^{(N)}_t\}$-adapted, θ is bounded, $\{\mathcal{F}^{(N)}_t\}$-adapted and left-continuous. Thus, using the fact that (5.21) is an intensity of $E^{(N),i}$, and by the identity

$$\begin{aligned} &\frac{1}{N}\varphi(a^{(N),i}(t-),t)\mathbb{1}_{\{X^{(N),i}(t-)=\ell-1\}} h^{(N)}_E(R^{(N)}_E(t-)) \\ &\qquad \times \sum_{\ell'=1}^{\infty} \mathbb{1}_{\{X^{(N),i}(t-)=\ell'-1\}} \mathfrak{P}_d(\overline{S}^{(N)}_{\ell'-1}(t-), \overline{S}^{(N)}_{\ell'}(t-)) \\ &\quad = \frac{1}{N}\varphi(a^{(N),i}(t-),t)\mathbb{1}_{\{X^{(N),i}(t-)=\ell-1\}} \\ &\qquad \times h^{(N)}_E(R^{(N)}_E(t-))\mathfrak{P}_d(\overline{S}^{(N)}_{\ell-1}(t-), \overline{S}^{(N)}_{\ell}(t-)), \end{aligned}$$

which holds since $\mathbb{1}_{\{X^{(N),i}(t-)=\ell-1\}}\mathbb{1}_{\{X^{(N),i}(t-)=\ell'-1\}} \neq 0$ if and only if $\ell' = \ell$, it follows from Lemma 5.4 that the process $\mathcal{N}^{(N),i}_{\varphi,\ell} := \mathcal{R}^{(N),i}_{\varphi,\ell} - \tilde{R}^{(N),i}_{\varphi,\ell}$, with

$$\begin{aligned} \tilde{R}^{(N),i}_{\varphi,\ell}(t) := \frac{1}{N}\int_0^t &\varphi(a^{(N),i}(s),s)\mathbb{1}_{\{X^{(N),i}(s)=\ell-1\}} h^{(N)}_E(R^{(N)}_E(s)) \\ &\times \mathfrak{P}_d(\overline{S}^{(N)}_{\ell-1}(s), \overline{S}^{(N)}_{\ell}(s))\, ds, \end{aligned}$$

is a local $\{\mathcal{F}^{(N)}_t\}$-martingale. Again, by (5.2) and (5.13),

$$\begin{aligned} \tilde{R}^{(N),i}_{\varphi,\ell}(t) &= \frac{1}{N}\int_0^t h^{(N)}_E(R^{(N)}_E(s))\mathfrak{P}_d(\overline{S}^{(N)}_{\ell-1}(s), \overline{S}^{(N)}_{\ell}(s)) \\ &\qquad \times \sum_{i=1}^{N} \varphi(a^{(N),i}(s),s)\mathbb{1}_{\{X^{(N),i}(s)=\ell-1\}}\, ds \\ &= B^{(N)}_{\varphi,\ell}(t). \end{aligned}$$

Using (5.30), it follows that $\sum_{i=1}^{N} \mathcal{N}_{\varphi,\ell}^{(N),i} = \mathcal{R}_{\varphi,\ell}^{(N)} - B_{\varphi,\ell}^{(N)} = \mathcal{N}_{\varphi,\ell}^{(N)}$, and hence, $\mathcal{N}_{\varphi,\ell}^{(N)}$ is also a local $\{\mathcal{F}_t^{(N)}\}$-martingale. Also, in the setup above, we have

$$\begin{aligned}\theta^2(s) &= \varphi^2\big(a^{(N),i}(s-), s\big)\mathbb{1}_{\{X^{(N),i}(s-)=\ell-1\}} \\ &= \big\langle \varphi^2(\cdot, s), \nu_{\ell-1}^{(N),i}(s-) - \nu_{\ell}^{(N),i}(s-)\big\rangle,\end{aligned}$$

and so (5.9) of Lemma 5.4 implies

$$[\mathcal{R}_{\varphi,\ell}^{(N),i}](t) = \int_0^t \big\langle \varphi^2(\cdot, s), \nu_{\ell-1}^{(N),i}(s-) - \nu_{\ell}^{(N),i}(s-)\big\rangle \, dE^{(N),i}(s) = \mathcal{R}_{\varphi^2,\ell}^{(N),i}.$$

Finally, by the same reasoning as in the case $\ell = 1$, for $\ell \geq 2$ we have $[\mathcal{N}_{\varphi,\ell}^{(N)}] = [\mathcal{R}_{\varphi,\ell}^{(N)}] = \mathcal{R}_{\varphi^2,\ell}^{(N)}$. This completes the proof. □

6. Proof of convergence results.

In Section 6.1, we prove relative compactness of the state and auxiliary processes. In Section 6.2, we first show that subsequential limits of the state processes satisfy the hydrodynamic equations, and then prove Theorem 2.6 and Corollary 2.8 in Section 6.2.3. For $H = \mathcal{D}_{\varphi,\ell}, \mathcal{A}_{\varphi,\ell}, \mathcal{M}_{\varphi,\ell}, \mathcal{R}_{\varphi,\ell}, \mathcal{B}_{\varphi,\ell}$, and $\mathcal{N}_{\varphi,\ell}$, let

$$\overline{H}^{(N)}(t) := \frac{H^{(N)}(t)}{N}, \qquad N \in \mathbb{N}, t \geq 0. \tag{6.1}$$

6.1. *Relative compactness.*

The relative compactness results are summarized in Theorem 6.16 of Section 6.1.4. They are established in Sections 6.1.2–6.1.4 by verifying the well-known criteria of Kurtz and Jakubowski summarized in Section 6.1.1.

6.1.1. *Review of relative compactness criteria.*

Recall from Section 1.4 that $w'(f, \cdot, \cdot)$ denotes the modulus of continuity of a function f in $\mathbb{D}_{\mathbb{R}}[0, \infty)$. The first result follows from Theorems 3.7.2, and and 3.8.6 and Remark 3.8.7 of [17].

PROPOSITION 6.1 (Kurtz's criteria). *A sequence of $\mathbb{R}$-valued càdlàg processes $\{Y^{(N)}\}_{N\in\mathbb{N}}$ is relatively compact if and only if $\{Y^{(N)}\}_{N\in\mathbb{N}}$ satisfies the following*:

K1. *For every rational $t \geq 0$,*

$$\lim_{r\to\infty} \sup_N \mathbb{P}\{|Y^{(N)}(t)| > r\} = 0; \tag{6.2}$$

K2a. *For every $\eta > 0$ and $T > 0$, there exists $\delta > 0$ such that*

$$\sup_N \mathbb{P}\{w'(Y^{(N)}, \delta, T) \geq \eta\} \leq \eta. \tag{6.3}$$

Moreover, $\{Y^{(N)}\}_{N\in\mathbb{N}}$ is relatively compact if it satisfies K1 *and the condition* K2b:

K2b. *For each $T \geq 0$, there exists $\beta > 0$ such that*

$$\lim_{\delta \to 0} \sup_N \mathbb{E}\Big[\sup_{0 \leq t \leq T} \big|Y^{(N)}(t+\delta) - Y^{(N)}(t)\big|^\beta\Big] = 0. \tag{6.4}$$

If $\mathcal{Z} = [0, L)$ or $\mathcal{Z} = [0, L) \times [0, \infty)$, equipped with the Euclidean metric, then $\mathbb{M}_F(\mathcal{Z})$ (defined in Section 1.4) is a separable metric space, and hence, a completely regular topological space with metrizable compacts. Thus, the next result follows from Theorem 4.6 of [21] and Theorem 5.1 in Chapter 1 of [7]. Recall that a family $\mathbb{F}$ of real continuous functionals on $\mathbb{M}_F(\mathcal{Z})$ is said to separate points in $\mathbb{M}_F(\mathcal{Z})$ if for every distinct $\mu, \tilde{\mu} \in \mathbb{M}_F(\mathcal{Z})$, there exists a function $f \in \mathbb{F}$ such that $f(\mu) \neq f(\tilde{\mu})$.

PROPOSITION 6.2 (Jakubowski's criteria). *For $\mathcal{Z} = [0, L)$ or $\mathcal{Z} = [0, L) \times [0, \infty)$, A sequence $\{\pi^{(N)}\}_{N \in \mathbb{N}}$ of $\mathbb{D}_{\mathbb{M}_F(\mathcal{Z})}[0, \infty)$-valued random elements is tight if and only if*:

J1. (*Compact containment condition.*) *For each $T > 0$ and $\eta > 0$, there exists a compact set $\mathcal{K}_{T,\eta} \subset \mathbb{M}_F(\mathcal{Z})$ such that*

$$\liminf_N \mathbb{P}\{\pi_t^{(N)} \in \mathcal{K}_{T,\eta} \text{ for all } t \in [0, T]\} > 1 - \eta.$$

J2. *There exists a family $\mathbb{F}$ of real continuous functionals on $\mathbb{M}_F(\mathcal{Z})$ that* (i) *is closed under addition, and* (ii) *separates points in $\mathbb{M}_F(\mathcal{Z})$, such that $\{\pi^{(N)}\}$ is $\mathbb{F}$-weakly tight, that is, for every $F \in \mathbb{F}$, the sequence $\{F(\pi_s^{(N)}); s \geq 0\}_{N \in \mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0, \infty)$.*

In particular, $\{\pi^{(N)}\}_{N \in \mathbb{N}}$ is relatively compact if it satisfies J1 *and* J2.

REMARK 6.3. A set that satisfies properties (i) and (ii) in condition J2 is

$$\mathbb{F} = \big\{F : \exists f \in \mathbb{C}_c^1(\mathcal{Z}) \text{ such that } F(\mu) = \langle f, \mu\rangle, \forall \mu \in \mathbb{M}_F(\mathcal{Z})\big\}.$$

6.1.2. *Relative compactness of sequences of departures and auxiliary processes.* First, note that using (4.19) and (2.18), we have for $\varphi \in \mathbb{C}_b([0, L) \times \mathbb{R}_+)$ and $\ell \geq 1$,

$$\limsup_{N \to \infty} \mathbb{E}\big[\overline{\mathcal{D}}_{\varphi^2,\ell}^{(N)}(t)\big] \leq \|\varphi\|_\infty^2 \limsup_{N \to \infty} \mathbb{E}\big[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(t)\big] < \infty. \tag{6.5}$$

LEMMA 6.4. *Suppose Assumptions* I, II(a) *and* III(a) *hold, and fix $\varphi \in \mathbb{C}_b([0, L) \times \mathbb{R}_+)$, $\ell \geq 1$. Then $\mathcal{M}_{\varphi,\ell}^{(N)}$ is a square integrable martingale. Moreover, for $t \geq 0$,*

$$\limsup_{N \to \infty} \mathbb{P}\Big\{\Big|\sup_{0 \leq s \leq t} \overline{\mathcal{M}}_{\varphi,\ell}^{(N)}(s)\Big| > \epsilon\Big\} = 0, \qquad \epsilon > 0, \tag{6.6}$$

and $\overline{\mathcal{M}}_{\varphi,\ell}^{(N)} \Rightarrow 0$ in $\mathbb{D}_{\mathbb{R}}[0, \infty)$, as $N \to \infty$.

PROOF. By Proposition 5.2, $\mathcal{M}^{(N)}_{\varphi,\ell}$ is a local martingale with $[\mathcal{M}^{(N)}_{\varphi,\ell}] = \mathcal{D}^{(N)}_{\varphi^2,\ell}$. By (6.5) and Theorem 7.35 of [25], $\mathcal{M}^{(N)}_{\varphi,\ell}$ is a square integrable $\{\mathcal{F}^{(N)}_t\}$-martingale and for all $t \geq 0$, $\mathbb{E}[(\mathcal{M}^{(N)}_{\varphi,\ell}(t))^2] = \mathbb{E}[\mathcal{D}^{(N)}_{\varphi^2,\ell}(t)]$. By Doob's inequality and (6.1), for every $T \geq 0$ and $\epsilon > 0$, $\mathbb{P}\{|\sup_{t\in[0,T]} \overline{\mathcal{M}}^{(N)}_{\varphi,\ell}(t)| > \epsilon\} \leq \mathbb{E}[\overline{\mathcal{D}}^{(N)}_{\varphi^2,\ell}(T)]/N\epsilon^2$. Together with (6.5), this implies (6.6), which in turn shows that $\overline{\mathcal{M}}^{(N)}_{\varphi,\ell}$ converges to zero in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ in probability, and hence, in distribution. □

To obtain further tightness results on the departure processes, we recall another many-server model, the so-called GI/GI/N queue, studied in [24]. In a GI/GI/N queue, arriving jobs choose an idle server at random if there exists one, or if all servers are busy, join a common queue and enter service in a FCFS manner when servers become free. Equivalently, one can view the GI/GI/N as a network of N parallel servers in which each server has its own queue and newly arriving jobs join the queue that has the least residual work (see Section XII.1 of [6]). A fluid limit for the GI/GI/N queue in the same regime (i.e., when $N \to \infty$ and the arrival rate is proportional to N) is obtained in [24]. Although the $SQ(d)$ and the GI/GI/N models have very different routing mechanisms (leading to completely different dynamics of the cumulative service entry process $K^{(N)}$) the total departure process $D^{(N)}_1$ and the measure-valued process $\nu^{(N)}_1$ that keeps track of ages of all jobs in service in the $SQ(d)$ model share some common properties with their counterparts (denoted in [24] by $D^{(N)}$ and $\nu^{(N)}$, resp.) in the GI/GI/N model. As a result, the very same techniques used in [24] to prove certain tightness estimates for $\{D^{(N)}\}_{N\in\mathbb{N}}$ and $\{\nu^{(N)}\}_{N\in\mathbb{N}}$ in the GI/GI/N model can be applied to also prove analogous tightness estimates for the particular processes $\{D^{(N)}_1\}_{N\in\mathbb{N}}$ and $\{\nu^{(N)}_1\}_{N\in\mathbb{N}}$ in the $SQ(d)$ model. Thus, the latter are summarized in Lemmas 6.5–6.9 below, with only references to the corresponding result for the GI/GI/N queue of which they are a routine adaptation, rather than full proofs. However, for completeness, a more detailed justification of these results is also provided in provided in Appendix D of an extended version of this paper [5].

The first result is used to prove Lemma 6.6 below and, along with Lemma 6.8, to prove relative compactness of the sequence of state processes in Proposition 6.15.

LEMMA 6.5. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, for* $\ell \geq 1$ *and* $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $\{\overline{A}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$, $\{\overline{D}^{(N)}_{\ell}\}_{N\in\mathbb{N}}$ *and* $\{\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ *are relatively compact in* $\mathbb{D}_{\mathbb{R}}[0,\infty)$.

PROOF. One can verify Kurtz's conditions for $\{\overline{A}^{(N)}_{\varphi,1}\}_{N\in\mathbb{N}}$, $\{\overline{D}^{(N)}_1\}_{N\in\mathbb{N}}$ and $\{\overline{\mathcal{D}}^{(N)}_{\varphi,1}\}_{N\in\mathbb{N}}$ by establishing analogous estimates as in Lemmas 5.7 and 5.8 of [24] using exactly the same techniques. Moreover, Kurtz's conditions are satisfied for

$\{\overline{A}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ for $\ell \geq 2$ because the fact that $\langle f, \nu_1^{(N)}\rangle \geq \langle f, \nu_\ell^{(N)}\rangle$ for every nonnegative function f, implies that for $0 \leq s \leq t$, we have $|A^{(N)}_{\varphi,\ell}(t)| \leq A^{(N)}_{|\varphi|,1}(t)$, and $|A^{(N)}_{\varphi,\ell}(t) - A^{(N)}_{\varphi,\ell}(s)| \leq |A^{(N)}_{|\varphi|,1}(t) - A^{(N)}_{|\varphi|,1}(s)|$. This proves the relative compactness of $\{\overline{A}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ for $\ell \geq 2$. By Lemma 6.4, the sequence $\{\overline{\mathcal{M}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ converges weakly to zero and thus is relatively compact, and hence, by (5.6), the sequence $\{\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}\}$ is also relatively compact. Setting $\varphi = \mathbf{1}$, this implies relative compactness of $\{\overline{D}^{(N)}_{\ell}\}_{N\in\mathbb{N}}$ for $\ell \geq 2$. □

It follows from (4.19) that the linear functional $\mathcal{D}^{(N)}_{\cdot,\ell}(t) : \varphi \mapsto \mathcal{D}^{(N)}_{\varphi,\ell}(t)$ can be identified with a random finite nonnegative (Radon) measure on $[0, L) \times \mathbb{R}_+$ (see Section 1.4) and $\mathcal{D}^{(N)}_{\cdot,\ell} = \{\mathcal{D}^{(N)}_{\cdot,\ell}(t), t \geq 0\}$ can be viewed as an $\mathbb{M}_F([0, L) \times \mathbb{R}_+)$-valued process.

LEMMA 6.6. *Suppose Assumptions I–III hold. Then, for $\ell \geq 1$, $\{\overline{\mathcal{D}}^{(N)}_{\cdot,\ell}\}$ is relatively compact in $\mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0, \infty)$.*

PROOF. For $\ell \geq 1$, Jakubowski condition J2 holds by Lemma 6.5. For $\ell = 1$, condition J1 can be verified using the exact same techniques as in the proof of Lemma 5.13 of [24]. Specifically, one can show that for $\eta > 0$ and $T \geq 0$, there exist a constant $B(\eta) < \infty$ and a sequence $\{m(n, \eta)\}$ with $m(n, \eta) \to L$ as $n \to \infty$ such that

$$\mathbb{P}\{\overline{\mathcal{D}}^{(N)}_{\cdot,1}(t) \notin \mathcal{K}_\eta \text{ for some } t \in [0, T]\} \leq \eta, \tag{6.7}$$

for the compact subset $\mathcal{K}_\eta \subset \mathbb{M}_F([0, L) \times \mathbb{R}_+)$ defined by

$$\mathcal{K}_\eta := \Big\{\mu \in \mathbb{M}_F\big([0, L\big) \times \mathbb{R}_+) : \\ \langle \mathbf{1}, \mu\rangle \leq B(\eta), \mu\big(\big(m(n, \eta), L\big) \times R_+\big) \leq \frac{1}{n} \ \forall n \in \mathbb{N}\Big\}.$$

Furthermore, for $\ell \geq 2$, by (4.18), for every nonnegative measurable function φ and $t \geq 0$, $\mathcal{D}^{(N)}_{\varphi,\ell}(t) \leq \mathcal{D}^{(N)}_{\varphi,1}(t)$. Thus, the bound (6.7) holds with the same compact set $\mathcal{K}_\eta$ when $\mathcal{D}^{(N)}_{\varphi,1}$ is replaced by $\mathcal{D}^{(N)}_{\varphi,\ell}$, and so condition J1 holds for $\{\overline{\mathcal{D}}^{(N)}_{\cdot,\ell}\}$. □

The next two results are on properties of $\overline{\nu}^{(N)}_1$. Lemma 6.7 is used to prove relative compactness of the routing sequence in Lemma 6.14, while Lemma 6.8 is used to prove relative compactness of $\{\overline{\nu}^{(N)}_\ell\}$ in Lemma 6.15.

LEMMA 6.7. *Suppose Assumption III holds. Then*

$$\lim_{m\to L} \sup_N \mathbb{E}[\overline{\nu}^{(N)}_1(0, (m, L))] = 0, \tag{6.8}$$

and if $L < \infty$,

$$\lim_{m \to L} \sup_N \mathbb{E}\left[\int_{[0,m)} \frac{\overline{G}(m)}{\overline{G}(x)} \overline{\nu}_1^{(N)}(0, dx)\right] = 0. \tag{6.9}$$

PROOF. The result follows from Assumption III utilizing the exact same techniques as those used in the proof of Lemma 5.12 of [24]. □

LEMMA 6.8. *If Assumption III holds, then $\{\overline{\nu}_1^{(N)}\}_{N \in \mathbb{N}}$ satisfies condition* J1 *with*

$$\mathcal{K}_{\eta,T} = \{\mu \in \mathbb{M}_F([0, L)) : \nu([m, L)) \leq \eta\}, \tag{6.10}$$

for some $m < L$, if $L < \infty$, and with

$$\mathcal{K}_{\eta,T} = \left\{\mu \in \mathbb{M}_F(\mathbb{R}_+) : \mu(r(n) + T, \infty) \leq \frac{1}{n} \textit{ for all } n < -\lceil \log \eta / \log 2 \rceil\right\} \tag{6.11}$$

for some sequence $r(n) \uparrow \infty$ if $L = \infty$.

PROOF. The proof of this result parallels the proof of Lemma 5.12 of [24]. □

The next result, along with Lemma 6.4, is used in Section 6.1.4 to identify the limit of the weighted departure processes (Proposition 6.17).

LEMMA 6.9. *Suppose Assumptions I–III hold. For every $\ell \geq 1$, suppose $\{\overline{\nu}_\ell^{(N)}\}_{N \in \mathbb{N}}$ converges almost surely to some $\nu_\ell \in \mathbb{D}_{M_F[0,L)}[0, \infty)$ along a subsequence. Then*

$$\limsup_{N \to \infty} \mathbb{E}\left[\left|\overline{A}_{\varphi,\ell}^{(N)}(t) - \int_0^t \langle \varphi(\cdot, s) h(\cdot), \nu_\ell(s)\rangle\, ds\right|\right] = 0. \tag{6.12}$$

PROOF. For $\ell = 1$, the result can be established following the techniques of the proofs of Lemmas 5.16 and 6.17 of [24]. It can be verified that the same techniques also apply for $\ell \geq 2$ because $\langle f, \nu_\ell^{(N)}(t)\rangle \leq \langle f, \nu_1^{(N)}(t)\rangle$, and $\langle f, \nu_\ell(t)\rangle \leq \langle f, \nu_1(t)\rangle$ for all $t \geq 0$, $\ell \geq 1$ and every nonnegative measurable function f; see the proof of Lemma 6.9 of [5] for more details. □

6.1.3. *Relative compactness of arrival and routing processes.*

We first establish relative compactness of the arrival process sequence.

LEMMA 6.10. *Suppose Assumption I holds. Then $\{\overline{E}^{(N)}\}_{N \in \mathbb{N}}$ is relatively compact in $\mathbb{D}_\mathbb{R}[0, \infty)$, and*

$$\limsup_{N \to \infty} \mathbb{E}[\overline{E}^{(N)}(t)^2] < \infty, \qquad t \geq 0. \tag{6.13}$$

PROOF. Since $\overline{E}^{(N)} \to \lambda\mathbf{Id}$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ by Lemma 2.5, $\{\overline{E}^{(N)}\}_{N\in\mathbb{N}}$ is relatively compact. Moreover, by Assumption I there exists a delayed renewal process $\tilde{E}$ such that $E^{(N)}(t) = \tilde{E}(Nt)$. Let $U_{\tilde{E}}$ be the renewal measure associated with the interarrival distribution $G_{\tilde{E}}$. Then basic calculations (see equation (2.3) of [16]) show that

$$\mathbb{E}[\widetilde{E}(t)^2] \le U_{\tilde{E}}(t) + \int_0^t U_{\tilde{E}}(t-s)\,dU_{\tilde{E}}(s) \le 2U_{\tilde{E}}(t)^2.$$

Hence, by the elementary renewal theorem (see, e.g., Proposition V.1.4 of [6]),

$$\begin{aligned}\limsup_{N\to\infty} \mathbb{E}[\overline{E}^{(N)}(t)^2] &= \limsup_{N\to\infty} \frac{1}{N^2}\mathbb{E}[\widetilde{E}(Nt)^2] \\ &\le 2t^2 \limsup_{N\to\infty}\left(\frac{U_{\tilde{E}}(Nt)}{Nt}\right)^2 = 2\lambda^2 t^2,\end{aligned}$$

which proves (6.13). □

Next, we focus on the sequence $\{\mathcal{R}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$. By (4.13), for $\ell \ge 1$ and $0 \le s \le t$,

$$|\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t) - \overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(s)| \le \|\varphi\|_\infty(\overline{E}^{(N)}(t) - \overline{E}^{(N)}(s)). \tag{6.14}$$

LEMMA 6.11. *Suppose Assumption I holds. For $\varphi \in \mathbb{C}_b([0,L)\times\mathbb{R}_+)$ and $\ell \ge 1$,*

$$\limsup_{N\to\infty} \mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t)] \le \|\varphi\|_\infty \limsup_{N\to\infty}\mathbb{E}[\overline{E}^{(N)}(t)] < \infty, \qquad t \ge 0, \tag{6.15}$$

and $\{\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ is relatively compact in $\mathbb{D}_{\mathbb{R}}[0,\infty)$.

PROOF. Inequality (6.15) follows from (6.14) with $s = 0$, and (2.17). Moreover, (6.14) shows that the modulus of continuity w' for $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ is bounded by that of $\overline{E}^{(N)}$. Relative compactness of $\{\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ then follows from that of $\{\overline{E}^{(N)}\}$ proved in Lemma 6.10, and the necessity and sufficiency of Kurtz's criteria K1 and K2a stated in Proposition 6.1. □

Next, recall the definitions of $B^{(N)}_{\varphi,\ell}$ and $\mathcal{N}^{(N)}_{\varphi,\ell}$ in (5.2) and (5.3), respectively.

LEMMA 6.12. *Suppose Assumptions I, II(a) and III(a) hold. Then, for $\varphi \in \mathbb{C}_b([0,L)\times\mathbb{R}_+)$ and $\ell \ge 1$, $\mathcal{N}^{(N)}_{\varphi,\ell}$ is a square integrable $\{\mathcal{F}^{(N)}_t\}$-martingale. Moreover, for $t \ge 0$,*

$$\limsup_{N\to\infty} \mathbb{P}\left\{\left|\sup_{0\le s\le t} \overline{\mathcal{N}}^{(N)}_{\varphi,\ell}(s)\right| > \epsilon\right\} = 0 \qquad \forall\epsilon > 0, \tag{6.16}$$

and $\overline{\mathcal{N}}^{(N)}_{\varphi,\ell} \Rightarrow 0$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$, as $N \to \infty$.

PROOF. By Proposition 5.1, $\mathcal{N}^{(N)}_{\varphi,\ell}$ is a local martingale with $[\mathcal{N}^{(N)}_{\varphi,\ell}] = \mathcal{R}^{(N)}_{\varphi^2,\ell}$. Since $\mathbb{E}[\mathcal{R}^{(N)}_{\varphi^2,\ell}(t)] < \infty$ by (6.15), by Theorem 7.35 of [25], $\mathcal{N}^{(N)}_{\varphi,\ell}$ is a square integrable martingale, and $\mathbb{E}[(\mathcal{N}^{(N)}_{\varphi,\ell}(t))^2] = \mathbb{E}[\mathcal{R}^{(N)}_{\varphi^2,\ell}(t)]$. By Doob's inequality, for $\epsilon > 0$, $\mathbb{P}\{|\sup_{t\in[0,T]}\overline{\mathcal{N}}^{(N)}_{\varphi,\ell}(t)| > \epsilon\} \le \mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\varphi^2,\ell}(T)]/N\epsilon^2$. Together with (6.15), this implies (6.16), which implies $\overline{\mathcal{N}}^{(N)}_{\varphi,\ell}$ converges to zero in distribution. □

The following result is a direct corollary of Lemmas 6.11 and 6.12.

COROLLARY 6.13. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, for* $\varphi \in \mathbb{C}_b([0,L)\times\mathbb{R}_+)$ *and* $\ell \ge 1$, $\{\overline{B}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ *is relatively compact in* $\mathbb{D}_{\mathbb{R}}[0,\infty)$.

Analogous to $\mathcal{D}^{(N)}_{\cdot,\ell}$, due to (6.14), the linear functional $\mathcal{R}^{(N)}_{\cdot,\ell} : \varphi \mapsto \mathcal{R}^{(N)}_{\varphi,\ell}$ can be identified with a random finite nonnegative measure on $[0,L)\times\mathbb{R}_+$, and hence, $\mathcal{R}^{(N)}_{\cdot,\ell} = \{\mathcal{R}^{(N)}_{\cdot,\ell}(t), t \ge 0\}$, can be viewed as an $\mathbb{M}_F([0,L)\times\mathbb{R}_+)$-valued process.

LEMMA 6.14. *Suppose Assumptions* I, II(a) *and* III *hold. Then, for every* $\ell \ge 1$, *the sequence* $\{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}\}_{N\in\mathbb{N}}$ *is relatively compact in* $\mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)$.

PROOF. By definition, $\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}$ is a pure jump process, and hence, lies in $\mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)$. Lemma 6.11 implies that for $\varphi \in \mathbb{C}_b([0,L)\times\mathbb{R}_+)$, $\{\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}\}_{N\in\mathbb{N}}$ is relatively compact (and, therefore, tight by Prohorov's theorem) in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. It only remains to show that $\{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}\}_{N\in\mathbb{N}}$ satisfies condition J1. We first claim that for $\ell \ge 1$,

$$\lim_{m\to L}\sup_N \mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T)] = 0. \tag{6.17}$$

Fix $m > 0$. Substituting $\varphi(x,s) = \mathbf{1}_{(m,L)}(x)$ in (4.14), we see that $\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m,L)},1} \equiv 0$. This proves (6.17) for $\ell = 1$. For $\ell \ge 2$, substituting $\varphi(x,s) = \mathbf{1}_{(m,L)}(x)$ in (4.47), we have

$$\begin{aligned}\mathcal{R}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T) = \sum_{j\ge j_0}\sum_{j'\ge 1} & \mathbb{1}_{\{\gamma^{(N)}_{j'}\le T\}}\mathbb{1}_{\{a^{(N)}_j(\gamma^{(N)}_{j'})>m\}}\mathbb{1}_{\{\alpha^{(N)}_j<\gamma^{(N)}_{j'}\le\beta^{(N)}_j\}} \\ & \times \mathbb{1}_{\{\chi^{(N)}_j(\gamma^{(N)}_{j'}-)=\ell-1\}}\mathbb{1}_{\{\kappa^{(N)}_{j'}=\kappa^{(N)}_j\}}.\end{aligned} \tag{6.18}$$

For every $j \ge j_0$, consider the (random) set

$$\begin{aligned}\mathcal{J}^{(N)}_{j,\ell}(T) := \{ & j' \ge 1 : \gamma^{(N)}_{j'} \le T, \gamma^{(N)}_{j'} \in [\alpha^{(N)}_j, \beta^{(N)}_j), \\ & \chi^{(N)}_j(\gamma^{(N)}_{j'}-) = \ell - 1, \kappa^{(N)}_{j'} = \kappa^{(N)}_j\},\end{aligned}$$

of jobs j' that arrive prior to T while job j is receiving service, and are routed to the same queue $\kappa_j^{(N)}$ as job j, and this queue has length $\ell - 1$ right before the arrival of j'. During the time $(\alpha_j^{(N)}, \beta_j^{(N)})$ when job j is in service at queue $\kappa_j^{(N)}$, there are no departures from the queue and so its length is nondecreasing. Thus, there can be at most one job j' that arrives during that period and finds the queue length equal to $\ell - 1$. In other words, when $\mathcal{J}_{j,\ell}^{(N)}(T) \neq \varnothing$, $\mathcal{J}_{j,\ell}^{(N)}(T) = \{j_*\}$ for some $j_* = j_*(j, \ell, N, T)$. Also, for $j > E^{(N)}(T)$, $\gamma_j^{(N)} > T$, and thus, $\mathcal{J}_{j,\ell}^{(N)}(T) = \varnothing$. Therefore, since all departure and arrival times are almost surely distinct (i.e., $P(\tilde{\Omega}_T) = 1$ by Corollary 5.6), almost surely,

$$\begin{aligned}\mathcal{R}_{\mathbf{1}_{(m,L)},\ell}^{(N)}(T) &= \sum_{j \geq j_0} \sum_{j' \in \mathcal{J}_{j,\ell}^{(N)}(T)} \mathbb{1}_{\{a_j^{(N)}(\gamma_{j'}^{(N)}) > m\}} \\ &= \sum_{j=j_0}^{E^{(N)}(T)} \mathbb{1}_{\{\mathcal{J}_{j,\ell}^{(N)}(T) \neq \varnothing\}} \mathbb{1}_{\{a_j^{(N)}(\gamma_{j_*}^{(N)}) > m\}}. \end{aligned} \tag{6.19}$$

Now we consider two possible cases. If $L = \infty$ and a job j has initial age $a_j^{(N)}(0) \leq m$, then $a_j^{(N)}(t) \leq m + T$ for all $t \in [0, T]$. Therefore, (6.19) implies

$$\mathbb{E}[\overline{\mathcal{R}}_{\mathbf{1}_{(m+T,L)},\ell}^{(N)}(T)] \leq \mathbb{E}\left[\frac{1}{N} \sum_{j \geq j_0} \mathbb{1}_{\{a_j^{(N)}(0) > m\}}\right] = \mathbb{E}[\overline{\nu}_1^{(N)}(0, (m, \infty))]. \tag{6.20}$$

On the other hand, if $L < \infty$, using (6.19) and the fact that the age process $a_j^{(N)}(\cdot)$ is nondecreasing and bounded by the service time v_j, we have

$$\begin{aligned}\mathcal{R}_{\mathbf{1}_{(m,L)},\ell}^{(N)}(T) \leq{}& \sum_{j=j_0}^{0} \mathbb{1}_{\{a_j^{(N)}(0) > m\}} + \sum_{j=j_0}^{0} \mathbb{1}_{\{a_j^{(N)}(0) \leq m\}} \mathbb{1}_{\{v_j > m\}} \\ &+ \sum_{j=1}^{E^{(N)}(T)} \mathbb{1}_{\{v_j > m\}}. \end{aligned} \tag{6.21}$$

Using the fact that j_0, $a_j^{(N)}(0)$, and $\nu_1^{(N)}(0)$ are $\mathcal{F}_0^{(N)}$-measurable, we then have

$$\begin{aligned}\mathbb{E}\left[\sum_{j=j_0}^{0} \mathbb{1}_{\{a_j^{(N)}(0) \leq m\}} \mathbb{1}_{\{v_j > m\}}\right] &= \mathbb{E}\left[\sum_{j=j_0}^{0} \mathbb{1}_{\{a_j^{(N)}(0) \leq m\}} \mathbb{E}[\mathbb{1}_{\{v_j > m\}} | \mathcal{F}_0^{(N)}]\right] \\ &= \mathbb{E}\left[\sum_{j=j_0}^{0} \mathbb{1}_{\{a_j^{(N)}(0) \leq m\}} \frac{\overline{G}(m)}{\overline{G}(a_j^{(N)}(0))}\right] \\ &= \mathbb{E}\left[\int_{[0,m)} \frac{\overline{G}(m)}{\overline{G}(x)} \nu_1^{(N)}(0, dx)\right]. \end{aligned}$$

Hence, taking expectations in (6.21) and using the independence of the initial conditions and the arrival process for the third term on the right-hand side, we have

$$\mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m,L)},\ell}(T)] \le \mathbb{E}[\overline{\nu}^{(N)}_1(0,(m,L))] + \mathbb{E}\biggl[\int_{[0,m)} \frac{\overline{G}(m)}{\overline{G}(x)}\overline{\nu}^{(N)}_1(0,dx)\biggr] \tag{6.22}$$
$$+ \overline{G}(m)\mathbb{E}[\overline{E}^{(N)}(T)].$$

Taking first the supremum over N and then the limit as $m \to L$ of both sides of (6.20) and (6.22), the claim (6.17) for both cases follows by using (6.8) and (6.9) in Lemma 6.7, the bound (2.17) and the elementary identity $\lim_{m\to L}\overline{G}(m) = 0$.

We now use the claim (6.17) to complete the proof of the lemma. Fix $\eta > 0$ and apply (6.15) with $\varphi = \mathbf{1}$ and $t = T$, to conclude that the constant

$$C(\eta,T) := \frac{2}{\eta}\sup_N \mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\mathbf{1},\ell}(T)] \tag{6.23}$$

is finite. Also, by (6.17), there exists a sequence $\{m(n,\eta)\}$ with $\lim_{n\to\infty} m(n,\eta) = L$ such that

$$\sup_N \mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m(n,\eta),L)},\ell}(T)] \le \frac{\eta}{n2^{n+1}}. \tag{6.24}$$

The subset $\mathcal{K}_{\eta,T}$ of $\mathbb{M}_F([0,L)\times\mathbb{R}_+)$ defined as

$$\mathcal{K}_{\eta,T} := \biggl\{\mu \in \mathbb{M}_F\bigl([0,L)\times\mathbb{R}_+\bigr) :$$
$$\langle \mathbf{1},\mu\rangle \le C(\eta,T), \mu\bigl((m(n,\eta),L)\times\mathbb{R}_+\bigr) \le \frac{1}{n}\ \forall n \in \mathbb{N}\biggr\},$$

is compact because $\sup_{\mu\in\mathcal{K}_{\eta,T}} \mu([0,L)\times\mathbb{R}_+) \le C(\eta,T)$ and

$$\inf_{\substack{C\subset[0,L)\times\mathbb{R}_+\\ C\text{ compact}}} \sup_{\mu\in\mathcal{K}_{\eta,T}} \mu(C^c) \le \inf_n \sup_{\mu\in\mathcal{K}_{\eta,T}} \mu\bigl((m(n,\eta),L)\times\mathbb{R}_+\bigr) = 0.$$

Finally, using (6.23) and (6.24), for $W := \{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}(t) \notin \mathcal{K}_\eta$ for some $t \in [0,T]\}$, we have

$$\begin{aligned}\mathbb{P}\{W\} &\le \mathbb{P}\{\overline{\mathcal{R}}^{(N)}_{\mathbf{1},\ell}(T) \ge C(\eta,T)\} + \sum_{n\in\mathbb{N}} \mathbb{P}\biggl\{\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m(n,\eta),L)},\ell}(T) > \frac{1}{n}\biggr\}\\ &\le \sup_N \frac{\mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\mathbf{1},\ell}(T)]}{C(\eta,T)} + \sum_{n\in\mathbb{N}} n\mathbb{E}[\overline{\mathcal{R}}^{(N)}_{\mathbf{1}_{(m(n,\eta),L)},\ell}(T)]\\ &\le \eta.\end{aligned}$$

This shows that $\{\overline{\mathcal{R}}^{(N)}_{\cdot,\ell}\}_{N\in\mathbb{N}}$ satisfies condition J1, and completes the proof. □

6.1.4. *Relative compactness of state variables.*

PROPOSITION 6.15. *Suppose Assumptions* I, II(a) *and* III(a) *hold. Then, for every* $\ell \geq 1$, *the sequence* $\{\overline{\nu}_\ell^{(N)}\}_{N\in\mathbb{N}}$ *is relatively compact in* $\mathbb{D}_{\mathbb{M}_F[0,L)}[0,\infty)$.

PROOF. By Lemma 6.7, the sequence $\{\overline{\nu}_1^{(N)}\}_{N\in\mathbb{N}}$ satisfies condition J1 with the compact set $\mathcal{K}_{\eta,T}$ equal to either (6.11) or (6.10) depending on whether L is finite or infinite. Moreover, for every interval I, $t \geq 0$ and $\ell \geq 2$, $\overline{\nu}_\ell^{(N)}(t, I) \leq \overline{\nu}_1^{(N)}(t, I)$, and hence, condition J1 also holds for $\{\overline{\nu}_\ell^{(N)}\}$ with the same set $\mathcal{K}_{\eta,T}$.

It remains to prove that $\{\nu_\ell^{(N)}\}$ satisfies condition J2, for which it suffices, by Remark 6.3, to show that for $f \in \mathbb{C}_c^1[0, L)$, $\{\langle f, \overline{\nu}_\ell^{(N)}(t)\rangle\}_{N\in\mathbb{N}}$ is tight in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. It follows from Proposition 4.5 and Remark 4.9 that

$$\begin{aligned}\langle f, \overline{\nu}_\ell^{(N)}(t)\rangle = \langle f, \overline{\nu}_\ell^{(N)}(0)\rangle + \int_0^t \langle f', \overline{\nu}_\ell^{(N)}(s)\rangle\, ds - \overline{\mathcal{D}}_{f,\ell}^{(N)}(t)\\ + f(0)\overline{D}_{\ell+1}^{(N)}(t) + \overline{\mathcal{R}}_{f,\ell}^{(N)}(t).\end{aligned}$$

Relative compactness of $\{\langle f, \overline{\nu}_\ell^{(N)}(0)\rangle\}_{N\in\mathbb{N}}$ follows from Assumption III(b), relative compactness of $\{\overline{D}_{\ell+1}^{(N)}\}_{N\in\mathbb{N}}$ and $\{\overline{\mathcal{D}}_{f,\ell}^{(N)}\}_{N\in\mathbb{N}}$ follow from Lemma 6.5 and relative compactness of $\{\overline{\mathcal{R}}_{f,\ell}^{(N)}\}_{N\in\mathbb{N}}$ follows from Lemma 6.11. Moreover, the fact that $\{\int_0^\cdot \langle f', \overline{\nu}_\ell^{(N)}(s)\rangle\, ds\}_{N\in\mathbb{N}}$ satisfies both criteria K1 and K2b, and hence, is relatively compact, follows from the bound $|\int_t^{t+\delta} \langle f', \overline{\nu}_\ell^{(N)}(s)\rangle\, ds| \leq \delta \|f'\|_\infty$, which uses the fact that $\overline{\nu}_\ell^{(N)}(s)$ is a subprobability measure. Therefore, for all $\ell \geq 1$ and $f \in \mathbb{C}_b^1[0,\infty)$, $\{\langle f, \overline{\nu}_\ell^{(N)}(t)\rangle\}_{N\in\mathbb{N}}$ is relatively compact in $\mathbb{D}_{\mathbb{R}}[0,\infty)$ (and, therefore, tight by Prohorov's theorem since $\mathbb{D}_{\mathbb{R}}[0,\infty)$ is Polish). □

We summarize all the results of this section in the following theorem.

THEOREM 6.16. *Suppose Assumptions* I, II(a) *and* III *hold. Then, for each* $\ell \geq 1$, *the sequence* $\{(\overline{\nu}^{(N)}, \overline{\mathcal{D}}_{\cdot,\ell}^{(N)}, \overline{\mathcal{R}}_{\cdot,\ell}^{(N)})\}_{N\in\mathbb{N}}$ *is relatively compact in* $\mathbb{D}_{\mathbb{S}}[0,\infty) \times \mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)^2$.

PROOF. By Proposition 6.15, for each $\ell \geq 1$, $\{\overline{\nu}_\ell^{(N)}\}_{N\in\mathbb{N}}$ is relatively compact in $\mathbb{M}_F[0, L)$. Therefore, by a diagonalization argument, every subsequence of $\{\overline{\nu}^{(N)}\}_{N\in\mathbb{N}} = \{(\overline{\nu}_\ell^{(N)}; \ell \geq 1)\}_{N\in\mathbb{N}}$ has a further subsequence that is convergent, simultaneously for all $\ell \geq 1$. This means that the sequence $\{\overline{\nu}^{(N)}\}_{N\in\mathbb{N}} \subset \mathbb{S}$ is relatively compact with respect to the metric $d_{\mathbb{S}}$ defined in (2.5). Relative compactness of the other two components are proved in Lemmas 6.6 and 6.14, respectively. □

6.2. *Characterization of subsequential limits.* We now show that any subsequential limit of the sequence $\{\nu^{(N)}\}_{N\in\mathbb{N}}$ is a solution to the hydrodynamic equations. Section 6.2.3 contains the proofs of Theorem 2.6, and Corollary 2.8.

6.2.1. *Limit of the departure processes.*

PROPOSITION 6.17. *Suppose Assumptions* I–III *hold, and fix* $\ell \geq 1$. *If* $(\overline{\mathcal{D}}^{(N)}_{\cdot,\ell}, \overline{\nu}^{(N)})$ *converges to* $(\mathcal{D}_{\cdot,\ell}, \nu)$ *in* $\mathbb{D}_{\mathbb{M}_F([0,L)\times[0,\infty))}[0,\infty) \times \mathbb{D}_{\mathbb{S}}[0,\infty)$ *almost surely, then, for every* $\varphi \in \mathbb{C}_b([0,L)\times\mathbb{R}_+)$, $\mathcal{D}_{\varphi,\ell}$ *is continuous and for every* $t \geq 0$,

$$\mathcal{D}_{\varphi,\ell}(t) = \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle\, ds, \qquad a.s. \tag{6.25}$$

PROOF. Using (5.6), the difference of the two sides of (6.25) is equal to

$$\mathcal{D}_{\varphi,\ell}(t) - \overline{\mathcal{D}}^{(N)}_{\varphi,\ell}(t) + \overline{\mathcal{M}}^{(N)}_{\varphi,\ell}(t) + \overline{A}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle\, ds. \tag{6.26}$$

By the hypothesis of this proposition, $\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}$ converges to $\mathcal{D}_{\varphi,\ell}$ almost surely, as $N \to \infty$. Moreover, by Lemma 5.5, only one departure can occur at each time, and hence, the jump size of $\overline{\mathcal{D}}^{(N)}_{\varphi,\ell}$ is bounded by $\|\varphi\|_\infty/\mathrm{N}$. Therefore, by Theorem 13.4 of [7], $\mathcal{D}_{\varphi,\ell}$ is continuous and the convergence also holds uniformly on compact sets, almost surely. Since $\overline{\mathcal{M}}^{(N)}_{\varphi,\ell}(t)$ also converges to zero in probability by Lemma 6.4, in view of (6.26) to prove (6.25) it suffices to show that

$$\limsup_{N\to\infty} \mathbb{E}\left[\left|\overline{A}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \nu_\ell(s)\rangle\, ds\right|\right] = 0. \tag{6.27}$$

However, (6.27) follows from Lemma 6.9(b). □

6.2.2. *Limit of the routing processes.* Analysis of the limit of the sequence of routing processes $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ is far more subtle than of the sequence of departure processes. This is because the backward recurrence time of the arrival process component $R^{(N)}_E$ of the state variable evolves at a faster time scale compared to the measure-valued process $\bar{\nu}^{(N)}$. Hence, computation of the limit of the sequence of compensators $\overline{B}^{(N)}_{\varphi,\ell}$ of the routing processes requires establishing a form of averaging principle for the fast and slow components. We begin with a preliminary result, which uses the bounds obtained in Lemma A.1.

LEMMA 6.18. *Suppose Assumption* I *holds. Then the following statements are true*:

1. *For every* $0 \leq s \leq t$, *the following limit holds in probability as* $N \to \infty$:

$$\frac{1}{N}\int_s^t h^{(N)}_E(R^{(N)}_E(u))\, du \to (t-s)\lambda. \tag{6.28}$$

2. *For every $t \geq 0$,*

$$C_E(t) := \limsup_{N\to\infty} \mathbb{E}\bigg[\bigg(\frac{1}{N}\int_0^t h_E^{(N)}(R_E^{(N)}(u))\,du\bigg)^2\bigg] < \infty. \tag{6.29}$$

PROOF. By Assumption I, $E^{(N)}$ is a pure renewal process with interarrival distribution $G_E^{(N)}$ and backward recurrence time $R_E^{(N)}$. Applying Lemma A.1 with P^*, G^*, h^*, r^* and r_0^* replaced by $E^{(N)}$, $G_E^{(N)}$, $h_E^{(N)}$, $R_E^{(N)}$ and $R^{(N)}$, respectively, it follows from (A.3) that for every $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}\bigg\{\bigg|\frac{1}{N}\int_0^t h_E^{(N)}(R_E^{(N)}(s))\,ds - \lambda t\bigg| > 2\epsilon\bigg\}\\
&\quad \leq \frac{1}{\epsilon^2}\mathbb{E}\bigg[\bigg(\frac{1}{N}\int_0^t h_E^{(N)}(R_E^{(N)}(s))\,ds - \overline{E}^{(N)}(t)\bigg)^2\bigg] + \mathbb{P}\{|\overline{E}^{(N)}(t) - \lambda t| > \epsilon\}\\
&\quad \leq \frac{3}{N\epsilon^2}\bigg(\frac{4}{N} + \mathbb{E}[\overline{E}^{(N)}(t)]\bigg) + \mathbb{P}\{|\overline{E}^{(N)}(t) - \lambda t| > \epsilon\}.
\end{aligned}$$

By (2.17), $\mathbb{E}[\overline{E}^{(N)}(t)]$ is finite, and $\overline{E}^{(N)}(t)$ converges in expectation, and hence, in probability, to λt by Lemma 6.10. Hence, the right-hand side of display above converges to zero as $N \to \infty$. Since both sides of (6.28) are linear in t, (6.28) follows.

To establish (6.29), by another application of (A.3) we have

$$\begin{aligned}
&\mathbb{E}\bigg[\bigg(\frac{1}{N}\int_0^t h_E^{(N)}(R_E^{(N)}(s))\,ds\bigg)^2\bigg]\\
&\qquad \leq \frac{2}{N^2}\mathbb{E}\bigg[\bigg(\int_0^t h_E^{(N)}(R_E^{(N)}(s))\,ds - E^{(N)}(t)\bigg)^2\bigg] + 2\mathbb{E}[\overline{E}^{(N)}(t)^2]\\
&\qquad \leq \frac{24}{N^2} + \frac{6}{N}\mathbb{E}[\overline{E}^{(N)}(t)] + 2\mathbb{E}[\overline{E}^{(N)}(t)^2].
\end{aligned}$$

Taking the limit superior as $N \to \infty$ of both sides of the inequality above and using (2.17), (6.29) follows with $C_E(t) = 2\limsup_{N\to\infty}\mathbb{E}[\overline{E}^{(N)}(t)^2]$, which is finite by (6.13). □

PROPOSITION 6.19. *Suppose Assumptions I, II(a) and III hold, fix $\ell \geq 1$, and let η_ℓ be defined as in (2.14). If $(\overline{\mathcal{R}}_{\cdot,\ell}^{(N)}, \overline{\nu}^{(N)})$ converges to $(\mathcal{R}_{\cdot,\ell}, \nu)$ in $\mathbb{D}_{\mathbb{M}_F([0,L)\times[0,\infty))}[0,\infty) \times \mathbb{D}_{\mathbb{S}}[0,\infty)$, almost surely as $N \to \infty$, then for every $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $\mathcal{R}_{\varphi,\ell}$ is continuous and satisfies for every $t \geq 0$,*

$$\mathcal{R}_{\varphi,\ell}(t) = \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s)\rangle\,ds, \qquad a.s. \tag{6.30}$$

PROOF. Fix $\varphi \in \mathbb{C}_b([0,L) \times \mathbb{R}_+)$, $\epsilon > 0$ and let $\mathcal{W} := \{|\mathcal{R}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s)\rangle\, ds| > 3\epsilon\}$. Using the relation $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell} = \overline{B}^{(N)}_{\varphi,\ell} + \overline{\mathcal{N}}^{(N)}_{\varphi,\ell}$ from (5.3), we have

$$\begin{aligned}\mathbb{P}\{\mathcal{W}\} &\le \mathbb{P}\{|\mathcal{R}_{\varphi,\ell}(t) - \overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t)| > \epsilon\} + \mathbb{P}\{|\overline{\mathcal{N}}^{(N)}_{\varphi,\ell}(t)| > \epsilon\}\\ &\quad + \mathbb{P}\left\{\left|\overline{B}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s)\rangle\, ds\right| > \epsilon\right\}.\end{aligned} \tag{6.31}$$

By the hypothesis of this proposition, $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ converges almost surely to $\mathcal{R}_{\varphi,\ell}$, in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Also, by (6.14), the jump sizes of $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ are bounded by $\|\varphi\|_\infty$ times the jump sizes of $\overline{E}^{(N)}$, which are at most $1/N$. Therefore, the maximum jump size of $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ converges to zero as $N \to \infty$, and by Theorem 13.4 of [7], $\mathcal{R}_{\varphi,\ell}$ is almost surely continuous and $\overline{\mathcal{R}}^{(N)}_{\varphi,\ell}$ converges to $\mathcal{R}_{\varphi,\ell}$, almost surely, uniformly on compact sets:

$$\limsup_{N\to\infty} \mathbb{P}\{|\mathcal{R}_{\varphi,\ell}(t) - \overline{\mathcal{R}}^{(N)}_{\varphi,\ell}(t)| > \epsilon\} = 0, \qquad t \ge 0. \tag{6.32}$$

Moreover, by (5.2) and (2.14), we can write

$$\begin{aligned}\overline{B}^{(N)}_{\varphi,\ell}(t) &- \int_0^t \langle \varphi(\cdot,s), \eta_\ell(s)\rangle\, ds\\ &= \int_0^t \left(\frac{1}{N} h_E^{(N)}(R_E^{(N)}(u)) - \lambda\right) f_{\varphi,\ell}(u)\, du\\ &\quad + \frac{1}{N}\int_0^t h_E^{(N)}(R_E^{(N)}(u))\big(f^{(N)}_{\varphi,\ell}(u) - f_{\varphi,\ell}(u)\big)\, du,\end{aligned} \tag{6.33}$$

with

$$f_{\varphi,\ell}(u) := \begin{cases} \varphi(0,u)(1 - S_1(u)^d) & \text{if } \ell = 1,\\ \langle \varphi(\cdot,u), \nu_{\ell-1}(u) - \nu_\ell(u)\rangle \mathfrak{P}_d(S_{\ell-1}(u), S_\ell(u)) & \text{if } \ell \ge 2,\end{cases}$$

and $f^{(N)}$ defined analogously, but with S_ℓ and ν_ℓ replaced by $\overline{S}^{(N)}_\ell$ and $\overline{\nu}^{(N)}$, respectively, for $\ell \ge 1$. By the hypothesis of this proposition, almost surely, $\overline{\nu}^{(N)}_\ell$ converges weakly to ν_ℓ, and hence, almost surely, $\overline{S}^{(N)}_\ell = \langle \mathbf{1}, \overline{\nu}^{(N)}_\ell\rangle$ and $\langle \varphi, \overline{\nu}^{(N)}_{\ell-1} - \overline{\nu}^{(N)}_\ell\rangle$ converge to $S_\ell = \langle \mathbf{1}, \nu_\ell\rangle$ and $\langle \varphi, \nu_{\ell-1} - \nu_\ell\rangle$, respectively. By Lemma 5.5, there is at most one arrival or departure at each time, and hence, the maximum jump sizes of $\overline{S}^{(N)}_\ell$ and $\langle \varphi, \overline{\nu}^{(N)}_{\ell-1} - \overline{\nu}^{(N)}_\ell\rangle$ are bounded by $1/N$ and $\|\varphi\|_\infty/N$, respectively. Therefore, by Theorem 13.4 of [7], S_ℓ and $\langle \varphi, \nu_{\ell-1} - \nu_\ell\rangle$ are continuous. Consequently, $f_{\varphi,\ell}$ is continuous, and $f^{(N)}_{\varphi,\ell}$ converges to $f_{\varphi,\ell}$ uniformly on compact sets, almost surely.

Fix $t \ge 0$ and $\delta > 0$, and let $w_{f_{\varphi,\ell}}(\delta,t)$ be the modulus of continuity of $f_{\varphi,\ell}$, and define $\mathcal{P} = \{0 = t_0 < t_1 < \cdots < t_n = t\}$ to be a partition of size $\delta = \max_j |t_j -$

$t_{j-1}|$. We have

$$\lim_{N\to\infty}\mathbb{P}\left\{\sum_{j=0}^{n-1} f_{\varphi,\ell}(t_j)\left|\frac{1}{N}\int_{t_j}^{t_{j+1}} h_E^{(N)}\big(R_E^{(N)}(u)\big)\,du-(t_{j+1}-t_j)\lambda\right|>\epsilon\right\}=0$$

(6.34)

and, by the Markov and Cauchy–Schwarz inequalities and (6.29),

$$\begin{aligned}
&\mathbb{P}\left\{w_{f_{\varphi,\ell}}(\delta,t)\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\left|\frac{1}{N}h_E^{(N)}\big(R_E^{(N)}(u)\big)-\lambda\right|du>\epsilon\right\}\\
&\quad\le\frac{1}{\epsilon}\mathbb{E}\left[w_{f_{\varphi,\ell}}(\delta,t)\left(\frac{1}{N}\int_0^t h_E^{(N)}\big(R_E^{(N)}(u)\big)\,du+\lambda t\right)\right]\\
&\quad\le\frac{\sqrt{2}}{\epsilon}\mathbb{E}\big[w^2_{f_{\varphi,\ell}}(\delta,t)\big]^{\frac{1}{2}}\left(\mathbb{E}\left[\left(\frac{1}{N}\int_0^t h_E^{(N)}\big(R_E^{(N)}(u)\big)\,du\right)^2\right]+\lambda^2t^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Taking the limit superior as $N\to\infty$ of both sides of the display above, using (6.29) and (6.34), we conclude that

$$\begin{aligned}
&\limsup_{N\to\infty}\mathbb{P}\left\{\left|\int_0^t\left(\frac{1}{N}h_E^{(N)}\big(R_E^{(N)}(u)\big)-\lambda\right)f_{\varphi,\ell}(u)\,du\right|>2\epsilon\right\}\\
&\quad\le\frac{C(t)}{\epsilon}\big(\mathbb{E}\big[w^2_{f_{\varphi,\ell}}(\delta,t)\big]\big)^{1/2},
\end{aligned}$$

where $C(t):=\sqrt{2}(C_E(t)+\lambda^2t^2)^{1/2}$. Since $f_{\varphi,\ell}$ is continuous on $[0,t]$, $w^2_{f_{\varphi,\ell}}(\delta,t)$ is bounded and converges almost surely to zero as $\delta\to0$. Sending $\delta\to0$ in the last display and applying the bounded convergence theorem, we see that

$$\limsup_{N\to\infty}\mathbb{P}\left\{\left|\int_0^t\left(\frac{1}{N}h_E^{(N)}\big(R_E^{(N)}(u)\big)-\lambda\right)f_{\varphi,\ell}(u)\,du\right|>2\epsilon\right\}=0. \tag{6.35}$$

Similarly, applying the Markov and Cauchy–Schwarz inequalities, we have

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{N}\int_0^t h_E^{(N)}\big(R_E^{(N)}(u)\big)\big(f_{\varphi,\ell}^{(N)}(u)-f_{\varphi,\ell}(u)\big)\,du\right|>\epsilon\right\}\\
&\quad\le\frac{1}{\epsilon}\mathbb{E}\Big[\sup_{u\in[0,t]}\big|f_{\varphi,\ell}^{(N)}(u)-f_{\varphi,\ell}(u)\big|^2\Big]^{\frac{1}{2}}\mathbb{E}\left[\left(\frac{1}{N}\int_0^t h_E^{(N)}\big(R_E^{(N)}(u)\big)\,du\right)^2\right]^{\frac{1}{2}}.
\end{aligned}$$

Together with (6.29), the almost-sure uniform convergence of $f_{\varphi,\ell}^{(N)}$ to the bounded function $f^{\varphi,\ell}$, and the bounded convergence theorem, this implies

$$\limsup_{N\to\infty}\mathbb{P}\left\{\left|\frac{1}{N}\int_0^t h_E^{(N)}\big(R_E^{(N)}(u)\big)\big(f_{\varphi,\ell}^{(N)}(u)-f_{\varphi,\ell}(u)\big)\,du\right|>\epsilon\right\}=0. \tag{6.36}$$

From (6.33), (6.35) and (6.36), it follows that

$$\limsup_{N\to\infty} \mathbb{P}\left\{\left|\overline{B}^{(N)}_{\varphi,\ell}(t) - \int_0^t \langle \varphi(\cdot, s), \eta_\ell(s)\rangle\, ds\right| > \epsilon\right\} = 0. \tag{6.37}$$

Finally, (6.30) follows on sending first $N \to \infty$ on the right-hand side of (6.31), next invoking (6.32), (6.16) of Lemma 6.12 and (6.37), and then sending $\epsilon \downarrow 0$. □

6.2.3. *Proofs of the convergence theorem and propagation of chaos.*

PROOF OF THEOREM 2.6. By Assumption III, Lemma 6.10 and Theorem 6.16, the sequence

$$Y^{(N)} := \left(\overline{\nu}^{(N)}(0), \overline{E}^{(N)}, \overline{\nu}^{(N)}, \overline{\mathcal{D}}^{(N)}_{\cdot,\ell}, \overline{\mathcal{R}}^{(N)}_{\cdot,\ell}; \ell \geq 1\right), \qquad N \in \mathbb{N}, \tag{6.38}$$

is relatively compact in

$$\begin{aligned} \mathcal{Y} := \mathbb{S} \times \mathbb{D}_{\mathbb{R}}[0,\infty) \times \mathbb{D}_{\mathbb{S}}[0,\infty) \times \mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)^{\mathbb{N}_0} \\ \times \mathbb{D}_{\mathbb{M}_F([0,L)\times\mathbb{R}_+)}[0,\infty)^{\mathbb{N}_0}. \end{aligned} \tag{6.39}$$

Therefore, for every subsequence $\{\mathcal{Y}^{N_k}\}$, there exists a further subsequence $\{N_{k_j}\}$, such that as $j \to \infty$, $Y^{N_{k_j}}$ converges in distribution to a random element

$$Y := \left(\nu(0), \lambda \mathbf{Id}, \nu, \mathcal{D}_{\cdot,\ell}, \mathcal{R}_{\cdot,\ell}; \ell \geq 1\right), \tag{6.40}$$

that takes values in $\mathcal{Y}$. It follows from the Skorokhod representation theorem that there exists a probability space that supports $\mathcal{Y}$-valued random elements $\tilde{\mathcal{Y}}_{N_{k_j}}$ and the $\mathcal{Y}$-valued random element $\tilde{Y}$, such that $\tilde{\mathcal{Y}}^{(N_{k_j})} \overset{d}{=} Y^{(N_{k_j})}$ for every j, $Y \overset{d}{=} \tilde{Y}$, and as $j \to \infty$, $\tilde{\mathcal{Y}}^{(N_{k_j})} \to \tilde{Y}$ almost surely in $\mathcal{Y}$. With a slight abuse of notation, since we are only interested in distributional properties, we denote the subsequence $\{N_{k_j}\}$ just as $\{N\}$ and identify $\tilde{Y}^{(N)}$ and $\tilde{Y}$ with $Y^{(N)}$ and Y, respectively. Using this convention, we have

$$Y^{(N)} \to Y \qquad \text{in } \mathcal{Y}, \text{ a.s.} \tag{6.41}$$

Now, we uniquely characterize the subsequential limit Y. Fix $\ell \geq 1$. For $f \in \mathbb{C}_b[0, L)$, by (6.41), $\langle f, \overline{\nu}^{(N)}_\ell\rangle$ converge almost surely to $\langle f, \nu_\ell\rangle$ in $\mathbb{D}_{\mathbb{R}}[0,\infty)$. Since for every $N \in \mathbb{N}$, the maximum jump size of $\langle f, \overline{\nu}^{(N)}_\ell\rangle$ is bounded by $\|f\|_\infty/N$ (due to Lemma 5.5) the limit $\langle f, \nu_\ell\rangle$ is continuous, and hence ν is a continuous $\mathbb{S}$-valued process, almost surely. Next, let $\mathcal{T}$ be a countable dense subset of $\mathbb{R}_+$ which contains 0 (say the diadic numbers). For $t \in \mathcal{T}$, it follows from Proposition 6.17, with $\varphi = \mathbf{1}$, that the limit $\mathcal{D}_{\mathbf{1},\ell}$ of $\overline{D}^{(N)}_\ell = \overline{\mathcal{D}}^{(N)}_{\mathbf{1},\ell}$ takes the form

$$\mathcal{D}_{\mathbf{1},\ell}(t) = \int_0^t \langle h, \nu_\ell(s)\rangle\, ds < \infty. \tag{6.42}$$

Therefore, sending $N \to \infty$ on both sides of (4.25) in Proposition 6.19, for every $t \in \mathcal{T}$, the identity

$$\langle \mathbf{1}, \nu_\ell(t)\rangle - \langle \mathbf{1}, \nu_\ell(0)\rangle = D_{\ell+1}(t) + \int_0^t \langle \mathbf{1}, \eta_\ell(s)\rangle\, ds - D_\ell(t), \tag{6.43}$$

holds almost surely, where η is defined by (2.14). Moreover, almost surely the relations (6.42) and (6.43) hold simultaneously for all $\ell \geq 1$ and $t \in \mathcal{T}$ because $\mathcal{T}$ is countable and, therefore, for all $t \geq 0$ since both sides are continuous by Propositions 6.17 and 6.19.

Furthermore, let $\mathcal{C}$ be a countable dense subset of $\mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$, and fix $\varphi \in \mathcal{C}$ and $t \in \mathcal{T}$. By Proposition 4.5, for every $\ell \geq 1$ and $N \in \mathbb{N}$, $Y^{(N)}$ satisfies the equation (4.24). Since φ, φ_x and φ_s are all bounded continuous functions, $\langle \varphi(\cdot, t), \overline{\nu}_\ell^{(N)}(t)\rangle$ and $\int_0^t \langle \varphi_s(\cdot, s) + \varphi_x(\cdot, s), \overline{\nu}_\ell^{(N)}(s)\rangle\, ds$ converge almost surely to $\langle \varphi(\cdot, t), \nu_\ell(t)\rangle$ and $\int_0^t \langle \varphi_s(\cdot, s) + \varphi_x(\cdot, s), \nu_\ell(s)\rangle\, ds$, respectively. Also, as we have already shown, $\overline{D}_{\ell+1}^{(N)}$ converges to $D_{\ell+1}$ almost surely in $\mathbb{D}_{\mathbb{R}}[0, \infty)$ and, therefore, the associated sequence of Stieltjes integrals $\int_{[0,t]} \varphi(0, s)\, d\overline{D}_{\ell+1}^{(N)}(s)$ converges almost surely to $\int_{[0,t]} \varphi(0, s)\, dD_{\ell+1}(s)$. Finally, by Proposition 6.17, the sequence $\overline{\mathcal{D}}_{\varphi,\ell}^{(N)}(t)$ converges to $\int_0^t \langle \varphi(\cdot, s)h(\cdot), \nu_\ell(s)\rangle\, ds$, and by Proposition 6.19, $\overline{\mathcal{R}}_{\varphi,\ell}^{(N)}(t)$ converge to $\int_0^t \langle \varphi(\cdot, s), \eta_\ell(s)\rangle\, ds$. Sending $N \to \infty$ on both sides of (4.24), we then have

$$\begin{aligned}\langle \varphi(\cdot, t), \nu_\ell(t)\rangle = {} & \langle \varphi(\cdot, 0), \nu_\ell(0)\rangle + \int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \nu_\ell(s)\rangle\, ds \\ & - \int_0^t \langle \varphi(\cdot, s)h(\cdot), \nu_\ell(s)\rangle\, ds + \int_0^t \varphi(0, s)\, dD_{\ell+1}(s) \\ & + \int_0^t \langle \varphi(\cdot, s), \eta_\ell(s)\rangle\, ds,\end{aligned} \tag{6.44}$$

almost surely. Equation (6.44) above holds with probability one, simultaneously for all $\ell \geq 1$, $\varphi \in \mathcal{C}$ and $t \in \mathcal{T}$, because both $\mathcal{C}$ and $\mathcal{T}$ are countable. Moreover, since both sides of the equation above are continuous functions of t, the identity holds simultaneously for all $t \geq 0$, and since ν_ℓ, $\mathcal{D}_{\cdot,\ell}$ and $\mathcal{R}_{\cdot,\ell}$ are finite Radon measures, the identity holds simultaneously for all $\varphi \in \mathbb{C}_c^{1,1}([0, L) \times \mathbb{R}_+)$ using the dominated convergence theorem.

Consequently, it follows from (6.42), (6.43) and (6.44) and Proposition 3.1 that ν is a solution to the hydrodynamic equations (2.10)–(2.14) associated to $(\lambda, \nu(0))$, which is proved to be unique in Theorem 2.4. This provides a unique characterization of all subsequential limits of $\{\overline{\nu}^{(N)}\}$ and completes the proof. □

PROOF OF COROLLARY 2.8. Since queues and servers are homogeneous and the routing algorithm is symmetric with respect to the queue indices, the queue lengths and age distributions remain exchangeable for all finite times $t \geq 0$. In particular, for any permutation $\pi : \{1, \dots, N\} \mapsto \{1, \dots, N\}$,

$$\big(X^{(N),i}(t); i = 1, \dots, N\big) \overset{d}{=} \big(X^{(N),\pi(i)}(t); i = 1, \dots, N\big). \tag{6.45}$$

Recall that $S_\ell^{(N)}(t)$ is the number of queues of length of at least ℓ, that is, with $X^{(N),i}(t) \geq 1$. Therefore,

$$\begin{aligned} \mathbb{E}[\overline{S}_\ell^{(N)}(t)] &= \frac{1}{N}\mathbb{E}\Bigg[\sum_{i=1}^N \mathbb{1}_{\{X^{(N),i}(t) \geq \ell\}}\Bigg] = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\{X^{(N),i}(t) \geq \ell\} \\ &= \mathbb{P}\{X^{(N),1}(t) \geq \ell\}, \end{aligned} \tag{6.46}$$

where the last equality is due to (6.45). By Theorem 2.6, and since the solution ν to the hydrodynamic equations is continuous, for every $t \geq 0$, $\overline{\nu}^{(N)}(t) \Rightarrow \nu(t)$ in $\mathbb{S}$. Hence, by the continuous mapping theorem, $\overline{S}_\ell^{(N)}(t) \Rightarrow S_\ell(t)$. Since $\sup_N \overline{S}_\ell^{(N)}(t)$ is bounded by 1, $\{\overline{S}_\ell^{(N)}\}$ is uniformly integrable and so the convergence also holds in expectation.

To prove the second claim, note that by (6.45),

$$\begin{aligned} &\mathbb{E}\Bigg[\prod_{m=1}^k \overline{S}_{\ell_m}^{(N)}(t)\Bigg] \\ &\quad = \frac{1}{N^k}\mathbb{E}\Bigg[\sum_{i_1=1}^N, \dots, \sum_{i_k=1}^N \mathbb{1}_{\{X^{(N),i_1}(t) \geq \ell_1\}}, \dots, \mathbb{1}_{\{X^{(N),i_k}(t) \geq \ell_k\}}\Bigg] \\ &\quad = \frac{1}{N^k}\sum_{i_1=1}^N \cdots \sum_{i_k=1}^N \mathbb{P}\{X^{(N),i_1}(t) \geq \ell_1, \dots, X^{(N),i_k}(t) \geq \ell_k\} \\ &\quad = \mathbb{P}\{X^{(N),1}(t) \geq \ell_1, \dots, X^{(N),k}(t) \geq \ell_k\}. \end{aligned} \tag{6.47}$$

Since $\overline{\nu}^{(N)}(t) \Rightarrow \nu(t)$ in $\mathbb{S}$, by the continuous mapping theorem, $\prod_{m=1}^n \overline{S}_{\ell_m}^{(N)}(t) \Rightarrow \prod_{m=1}^n S_{\ell_m}(t)$ and, since $\sup_N \prod_{m=1}^n \overline{S}_{\ell_m}^{(N)}(t)$ is bounded by 1, the convergence also holds in expectation. Taking the limit as $N \to \infty$ of both sides of (6.47), (2.20) follows. □

APPENDIX: A BOUND FOR RENEWAL PROCESSES

Fix $r_0^* \in \mathbb{R}_+$ and let $P^*(t)$ be a delayed renewal process with interarrival times $\{u_n^*; n \geq 1\}$ with common distribution G^* and delay u_0^* with distribution $G_{r_0^*}^*$:

$$\mathbb{P}\{u_0^* \leq x\} = G_{r_0^*}^*(x) := \frac{G^*(x + r_0^*) - G^*(r_0^*)}{1 - G^*(r_0^*)}. \tag{A.1}$$

Assume G^* has a density, denote $\overline{G^*} := 1 - G^*$ and let h^* be the corresponding rate function. Also, let $r^*(t)$ denote the backward recurrence time of the renewal process P^*. By convention, $r^*(t) = r_0^* + t$ for $t < u_0^*$, and in particular, $r^*(0) = r_0^*$.

LEMMA A.1. *Given the quantities described above, for every $t \geq 0$,*

$$\mathbb{E}\biggl[\int_0^t h^*\bigl(r^*(s)\bigr)\,ds\biggr] < \infty \tag{A.2}$$

and

$$\mathbb{E}\biggl[\biggl(\int_0^t h^*\bigl(r^*(s)\bigr)\,ds - P^*(t)\biggr)^2\biggr] \leq 12 + 3\mathbb{E}\bigl[P^*(t)\bigr]. \tag{A.3}$$

PROOF. Define the epoch times $\{t_j; j \geq 0\}$ as $t_0 = u_0^*$ and $t_j = t_{j-1} + u_j^*$ for $j \geq 1$. Then we have

$$\begin{aligned}\int_0^t h^*\bigl(r^*(s)\bigr)\,ds &= \int_0^{t_0} h^*\bigl(r^*(s)\bigr)\,ds + \sum_{n=1}^{P^*(t)} \int_{t_{n-1}}^{t_n} h^*\bigl(r^*(s)\bigr)\,ds \\ &\quad - \int_t^{t_{P^*(t)}} h^*\bigl(r^*(s)\bigr)\,ds \\ &= \int_{r_0^*}^{r_0^*+u_0^*} h^*(v)\,dv + \sum_{n=1}^{P^*(t)} \int_0^{u_n^*} h^*(v)\,dv - \int_{r^*(t)}^{u^*_{P^*(t)}} h^*(v)\,dv. \end{aligned} \tag{A.4}$$

Defining the random variables $y_n, n \in \mathbb{N}, s$ as $y_n := \int_0^{u_n^*} h^*(v)\,dv$, the second term on the right-hand side of (A.4) can be written as $\sum_{n=1}^{P^*(t)} y_n$. Since the renewal times $\{u_n^*; n \geq 1\}$ are i.i.d. the sequence $\{y_n\}_{n\in\mathbb{N}}$ is also i.i.d. with

$$\mathbb{E}[y_1] = \int_0^\infty \biggl(\int_0^s h^*(v)\,dv\biggr) G^*(s)\,ds = \int_0^\infty \frac{G^*(s)}{\overline{G^*}(s)} \biggl(\int_s^\infty G^*(v)\,dv\biggr)\,ds = 1,$$

and $\mathbb{E}[(y_1)^2]$ is equal to

$$\begin{aligned}\int_0^\infty \biggl(\int_0^s h^*(v)\,dv\biggr)^2 G^*(s)\,ds &= \int_0^\infty \bigl(\log(\overline{G^*}(s))\bigr)^2 G^*(s)\,ds \\ &= \int_0^1 \bigl(\log(s)\bigr)^2\,ds = 2.\end{aligned}$$

Thus, the mean and variance of y_n are both equal to 1. Now, define the discrete-time filtration $\{\mathcal{G}_n; n \geq 0\}$ by $\mathcal{G}_n = \sigma(u_j^*; j = 0, \dots, n)$. Note that u_n^*, and hence, y_n, are $\mathcal{G}_n$-measurable. Also, since the interarrival times are independent, $y_{n+1}, y_{n+2}, \dots$ are independent of $\mathcal{G}_n$. Finally, the random variable $P^*(t)$ satisfies $\mathbb{E}[P^*(t)] \leq U^*(t) < \infty$. where U^* is the renewal measure corresponding to the distribution G^*. The inequality can be replaced by an equality if P^* is replaced with a pure renewal process $\tilde{P}^*$; see Theorem 2.4.(iii) in Section V of [6], and P^* and $\tilde{P}^*$ can be coupled such that almost surely, $P^*(t) \le \tilde{P}^*(t)$ for all $t \ge 0$. Moreover, $P^*(t)$ is an integrable $\{\mathcal{G}_n\}$-stopping time because $\{P^*(t) = n\} = \{t_{n-1} \le t < t_n\} \in \mathcal{G}_n$ since both t_{n-1} and t_n are $\mathcal{G}_n$-measurable. Hence, by Wald's lemma (see Proposition A.10.2 of [6]),

$$\mathbb{E}\left[\sum_{n=1}^{P^*(t)} \int_0^{u_n^*} h^*(v)\,dv\right] = \mathbb{E}\left[\sum_{n=1}^{P^*(t)} y_n\right] = \mathbb{E}\bigl[P^*(t)\bigr]\mathbb{E}[y_1] = \mathbb{E}\bigl[P^*(t)\bigr] < \infty, \tag{A.5}$$

and $\mathbb{E}[(\sum_{n=1}^{P^*(t)} \int_0^{u_n^*} h^*(v)\,dv - P^*(t))^2]$ is equal to

$$\mathbb{E}\left[\left(\sum_{n=1}^{P^*(t)} y_n - P^*(t)\right)^2\right] = \mathbb{E}\bigl[P^*(t)\bigr]\operatorname{Var}(y_1) = \mathbb{E}\bigl[P^*(t)\bigr]. \tag{A.6}$$

Now, for the first term on the right-hand side of (A.4), using (A.1), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\int_{r_0^*}^{r_0^*+u_0^*} h^*(v)\,dv\right]\\
&\quad = \frac{1}{\overline{G^*}(r_0^*)}\int_0^\infty \left(\int_{r_0^*}^{r_0^*+u} h^*(v)\,dv\right) G^*(r_0^*+u)\,du\\
&\quad = \frac{1}{\overline{G^*}(r_0^*)}\int_0^\infty (\log(\overline{G^*}(r_0^*)) - \log(\overline{G^*}(r_0^*+u)))G^*(r_0^*+u)\,du\\
&\quad = \frac{\log(\overline{G^*}(r_0^*))}{\overline{G^*}(r_0^*)}\int_{r_0^*}^\infty G^*(u)\,du - \frac{1}{\overline{G^*}(r_0^*)}\int_0^{\overline{G^*}(r_0^*)} \log(v)\,dv\\
&\quad = 1, \qquad\qquad \text{(A.7)}
\end{aligned}$$

and $\mathbb{E}[(\int_{r_0^*}^{r_0^*+u_0^*} h^*(v)\,dv)^2]$ is equal to

$$\begin{aligned}
&\frac{1}{\overline{G^*}(r_0^*)}\int_0^\infty \left(\int_{r_0^*}^{r_0^*+u} h^*(v)\,dv\right)^2 G^*(r_0^*+u)\,du\\
&\quad = \frac{1}{\overline{G^*}(r_0^*)}\int_0^\infty (\log(\overline{G^*}(r_0^*)) - \log(\overline{G^*}(r_0^*+u)))^2 G^*(r_0^*+u)\,du\\
&\quad = \log(\overline{G^*}(r_0^*))^2 - 2\frac{\log(\overline{G^*}(r_0^*))}{\overline{G^*}(r_0^*)}\int_0^{\overline{G^*}(r_0^*)} \log(v)\,dv\\
&\qquad + \frac{1}{\overline{G^*}(r_0^*)}\int_0^{\overline{G^*}(r_0^*)} (\log(v))^2\,dv\\
&\quad = \log(\overline{G^*}(r_0^*))^2 - 2\log(\overline{G^*}(r_0^*))(\log(\overline{G^*}(r_0^*)) - 1)
\end{aligned}$$

$$+\left(\log\left(\overline{G^*}(r_0^*)\right)^2 - 2\log\left(\overline{G^*}(r_0^*)\right) + 2\right)$$

$$= 2. \tag{A.8}$$

For the last term on the right-hand side of (A.4), since $r^*(t) \geq 0$,

$$\mathbb{E}\left[\int_{r^*(t)}^{u^*_{P^*(t)}} h^*(v)\,dv\right] \leq \mathbb{E}\left[\int_0^{u^*_{P^*(t)}} h^*(v)\,dv\right] = \mathbb{E}[y_{P^*(t)}] = 1 \tag{A.9}$$

and

$$\mathbb{E}\left[\left(\int_{r^*(t)}^{u^*_{P^*(t)}} h^*(v)\,dv\right)^2\right] \leq \mathbb{E}\left[\left(\int_0^{u^*_{P^*(t)}} h^*(v)\,dv\right)^2\right]$$

$$= \mathbb{E}\left[(y_{P^*(t)})^2\right] = 2. \tag{A.10}$$

Then (A.5) follows on taking expectations of both sides of (A.4) and using (A.5), (A.7) and (A.9), while (A.3) follows on again applying (A.4), the elementary bound $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, and invoking (A.6), (A.8) and (A.10). □

Acknowledgments. Both authors would like to thank Microsoft Research New England for their hospitality during the Fall of 2014 when part of this research was conducted.

REFERENCES

[1] AGARWAL, P. and RAMANAN, K. (2018). Invariant states of the hydrodynamic limit of the SQ(d) model with general service times. Preprint.

[2] AGHAJANI, R. (2017). Infinite-dimensional scaling limits of stochastic networks Ph.D. thesis, Brown Univ., Providence, RI.

[3] AGHAJANI, R., LI, X. and RAMANAN, K. (2015). Mean-field dynamics of load-balancing networks with general service distributions. Available at arXiv:1512.05056 [math.PR].

[4] AGHAJANI, R., LI, X. and RAMANAN, K. (2017). The PDE method for the analysis of randomized load balancing networks. *Proc. ACM Meas. Anal. Comput. Syst.* **1** 38:1–38:28.

[5] AGHAJANI, R. and RAMANAN, K. (2017). The hydrodynamic limit of a randomized load balancing network. Extended version. Available at arXiv:1707.02005 [math.PR].

[6] ASMUSSEN, S. (2003). *Applied Probability and Queues*: *Stochastic Modelling and Applied Probability*, 2nd ed. *Applications of Mathematics* (*New York*) **51**. Springer, New York. MR1978607

[7] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. MR0233396

[8] BRAMSON, M. (2011). Stability of join the shortest queue networks. *Ann. Appl. Probab.* **21** 1568–1625. MR2857457

[9] BRAMSON, M., LU, Y. and PRABHAKAR, B. (2010). Randomized load balancing with general service time distributions. *SIGMETRICS Perform. Eval. Rev.* **38** 275–286.

[10] BRAMSON, M., LU, Y. and PRABHAKAR, B. (2012). Asymptotic independence of queues under randomized load balancing. *Queueing Syst.* **71** 247–292. MR2943660

[11] BRAMSON, M., LU, Y. and PRABHAKAR, B. (2013). Decay of tails at equilibrium for FIFO join the shortest queue networks. *Ann. Appl. Probab.* **23** 1841–1878. MR3114919

[12] BRÉMAUD, P. (1981). *Point Processes and Queues*: *Martingale Dynamics*. *Springer Series in Statistics*. Springer, New York. MR0636252

[13] BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., SHEN, H., ZELTYN, S. and ZHAO, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50. MR2166068

[14] CHEN, H. and YAO, D. D. (2001). *Fundamentals of Queueing Networks*: *Performance*, *Asymptotics*, *and Optimization*, *Stochastic Modelling and Applied Probability*. *Applications of Mathematics* (*New York*) **46**. Springer, New York. MR1835969

[15] CHEN, S., SUN, Y., KOZAT, U. C., HUANG, L., SINHA, P., LIANG, G., LIU, X. and SHROFF, N. B. (2014). When queueing meets coding: Optimal-latency data retrieving scheme in storage clouds. In *INFOCOM* 2014—*IEEE Conference on Computer Communications* 1042–1050.

[16] DALEY, D. J. and MOHAN, N. R. (1978). Asymptotic behaviour of the variance of renewal processes and random walks. *Ann. Probab.* **6** 516–521. MR0474534

[17] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0838085

[18] FARIAS, V. F., MOALLEMI, C. C. and PRABHAKAR, B. (2005). Load balancing with migration penalties. In *Proceedings of the IEEE International Symposium on Information Theory* 558–562.

[19] GANESH, A., LILIENTHAL, S., MANJUNATH, D., PROUTIERE, A. and SIMATOS, F. (2012). Load balancing via random local search in closed and open systems. *Queueing Syst.* **71** 321–345. MR2943662

[20] GRAHAM, C. (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *J. Appl. Probab.* **37** 198–211. MR1761670

[21] JAKUBOWSKI, A. (1986). On the Skorokhod topology. *Ann. Inst. Henri Poincaré Probab. Stat.* **22** 263–285. MR0871083

[22] KANG, W. and RAMANAN, K. (2010). Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* **20** 2204–2260. MR2759733

[23] KARDASSAKIS, K. (2014). Load balancing in stochastic networks: Algorithms, analysis, and game theory. Undergraduate Honors Thesis, Brown Univ., Providence, RI.

[24] KASPI, H. and RAMANAN, K. (2011). Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* **21** 33–114. MR2759196

[25] KLEBANER, F. C. (2005). *Introduction to Stochastic Calculus with Applications*, 2nd ed. Imperial College Press, London. MR2160228

[26] KOLESAR, P. (1984). Stalking the endangered CAT: A queueing analysis of congestion at Automatic Teller Machines. *Interfaces* **14** 16–26.

[27] LIANG, G. and KOZAT, U. C. (2014). TOFEC: achieving optimal throughput-delay trade-off of cloud storage using erasure codes. In *INFOCOM* 2014—*IEEE Conference on Computer Communications* 826–834.

[28] LUCZAK, M. J. and NORRIS, J. (2005). Strong approximation for the supermarket model. *Ann. Appl. Probab.* **15** 2038–2061. MR2152252

[29] MITZENMACHER, M. (2001). Analyses of load stealing models based on families of differential equations. *Theory Comput. Syst.* **34** 77–98. MR1799068

[30] MITZENMACHER, M. (2001). The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.* **12** 1094–1104.

[31] MUKHOPADHYAY, A. and MAZUMDAR, R. R. (2016). Analysis of randomized join-the-shortest-queue (JSQ) schemes in large heterogeneous processor-sharing systems. *IEEE Trans. Control Netw. Syst.* **3** 116–126. MR3514587

[32] ROGERS, L. C. G. and WILLIAMS, D. (2000). *Diffusions*, *Markov Processes*, *and Martingales*. *Vol*. 2. *Itô Calculus*. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. Reprint of the second (1994) edition. MR1780932

[33] VVEDENSKAYA, N. D., DOBRUSHIN, R. L. and KARPELEVICH, F. I. (1996). A queueing system with a choice of the shorter of two queues—an asymptotic approach. *Problemy Peredachi Informatsii* **32** 20–34. MR1384927

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
9500 GILMAN DR.
LA JOLLA, CALIFORNIA 92093
USA
E-MAIL: maghajani@ucsd.edu

DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912
USA
E-MAIL: kavita_ramanan@brown.edu