The Annals of Applied Probability
2011, Vol. 21, No. 1, 33–114
DOI: 10.1214/09-AAP662

LAW OF LARGE NUMBERS LIMITS FOR MANY-SERVER QUEUES

BY HAYA KASPI[1,2] AND KAVITA RAMANAN[1,3]

Technion—Israel Institute of Technology and Brown University

This work considers a many-server queueing system in which customers with independent and identically distributed service times, chosen from a general distribution, enter service in the order of arrival. The dynamics of the system are represented in terms of a process that describes the total number of customers in the system, as well as a measure-valued process that keeps track of the ages of customers in service. Under mild assumptions on the service time distribution, as the number of servers goes to infinity, a law of large numbers (or fluid) limit is established for this pair of processes. The limit is characterized as the unique solution to a coupled pair of integral equations which admits a fairly explicit representation. As a corollary, the fluid limits of several other functionals of interest, such as the waiting time, are also obtained. Furthermore, when the arrival process is time-homogeneous, the measure-valued component of the fluid limit is shown to converge to its equilibrium. Along the way, some results of independent interest are obtained, including a continuous mapping result and a maximality property of the fluid limit. A motivation for studying these systems is that they arise as models of computer data systems and call centers.

CONTENTS

Received July 2007; revised November 2009.

[1]Supported in part by the US–Israel Binational Science Foundation under Grant BSF-2006379.

[2]Supported in part by the Technion Fund for the Promotion of Research under Grant 2003941 and the Milford Bohm Chair grant.

[3]Supported in part by NSF Grants DMS-04-06191, CMMI-1059967 (formerly CMMI-0728064) and CMMI-1052750 (formerly CMMI-0928154).

AMS 2000 subject classifications. Primary 60F17, 60K25, 90B22; secondary 60H99, 35D99.

Key words and phrases. Multi-server queues, $GI/G/N$ queue, fluid limits, mean-field limits, strong law of large numbers, measure-valued processes, call centers.

1. Introduction.

1.1. *Background and motivation*. The main objective of this work is to obtain functional strong laws of large numbers limits or "fluid" approximations of various functionals of the $G/GI/N$ queue (in other words, a queue with N servers that has a general cumulative arrival process, and in which customers with independent and identically distributed service times are processed in their order of arrival), in the limit as the number of servers tends to infinity. In fact, a more general setting is considered that allows for possibly time-inhomogeneous arrivals. In order to obtain a Markovian description of the dynamics, the state includes a nonnegative integer-valued process that represents the total number of customers in system, as well as a measure-valued process that keeps track of the ages of customers in service. The fluid limit obtained is for this pair of processes and thus contains more information than just the limit of the scaled number of customers in system. In particular, it also yields a description of the fluid limits of several other functionals of interest, including the waiting time. A fairly explicit representation for the fluid limit is obtained, which is then used to study the convergence, as $t \to \infty$, of the fluid limit to equilibrium in the case when the arrival process is time-homogeneous. These results are obtained under mild assumptions on the service distribution, such as the existence of a density, which are satisfied by most distributions that arise in applications. While we expect that these conditions can be relaxed, the representation of the fluid limit is likely to be more involved in that setting. Thus, for ease of exposition, we have restricted ourselves to this generality.

Multiserver queueing systems arise in many applications, and have generally proved to be more difficult to analyse than single server queues. Thus, it is natural to resort to an asymptotic analysis in order to gain insight into the behavior of these systems. It is of particular interest to consider an asymptotic regime in which the probability of a positive queue lies strictly between zero and one since this captures what is observed in many applications. In the seminal paper of Halfin and Whitt [8], it was shown that for the case of Poisson arrivals and exponentially distributed service times, this can be achieved by letting both the number of servers N and the corresponding arrival rate λ_N go to infinity in such a manner that $\lambda_N = N - \beta\sqrt{N}$, for some $\beta > 0$. Specifically, Halfin and Whitt [8] established a central limit theorem for the number of customers in system in this setting, and then used it to derive an approximation for the probability of a positive queue (equivalently, the probability of a customer having to wait a positive amount of time). For networks of multi-server queues with (possibly time-varying) Poisson arrivals and exponential services, fluid and diffusion limits for the total number of customers in system were obtained by Mandelbaum, Massey and Reiman [15]. All of these results were obtained under the assumption of exponential service times.

This work is to a large extent motivated by the fact that $G/GI/N$ queues arise as models of large-scale telephone call centers, for which the limiting regime considered here admits the natural interpretation of the scaling up of the number of

servers in response to an analogous scaling up of the arrival rate of customers (see Brown et al. [2] for a survey of applications of multi-server models to call centers). Recent statistical evidence from real call centers presented in Brown et al. [2] suggests that, in many cases, it may be more appropriate to model the service times as being nonexponentially distributed and, in particular, chosen from a lognormal distribution. This emphasizes the need to characterize fluid limits when the service times are generally distributed. With this as motivation, a deterministic fluid approximation for a $G/GI/N$ queue, with abandonments that are possibly generally distributed, was proposed by Whitt [22]. However, a general functional strong law of large numbers justifying the fluid approximation was not obtained in [22] (instead, convergence was established for a discrete-time version of the model, allowing for time-dependent and state-dependent arrivals). In this paper, we establish a functional strong law of large numbers limit allowing for time-dependent arrivals, but in the absence of abandonments, and provide an intuitive and fairly explicit characterization of the limit. We consider an asymptotic regime in which the number of servers N goes to infinity, and the arrival rate λ_N scales roughly as $N\lambda(\cdot)$, for some possibly time-varying function λ. This asymptotic regime is essentially the same as the one considered in [22], and is a slight generalization of the one introduced originally by Halfin and Whitt in [8], adapted to the situation in which only a "fluid" approximation (as opposed to a central limit theorem) is sought. Concurrently with this work, fluid and central limit theorems for just the number in system were established in the work of Reed [19] using a clever comparison with a $G/GI/\infty$ system. Here, we take a different approach that involves a measure-valued representation. This leads to a fairly explicit representation of the fluid limit of the number in system and also yields the fluid limits of several other functionals of interest.

One of the challenges in going from exponential to nonexponential service distributions is that a Markovian description of the dynamics leads, in the limit as $N \to \infty$, to an infinite-dimensional state. The measure-valued representation and martingale methods adopted in this paper provide a convenient framework for the asymptotic analysis of multi-server queues (see Pang et al. [17] for a recent survey on the use of martingale methods for establishing heavy-traffic limits of multi-server queues with exponentially distributed service times). Indeed, the framework developed here is quite flexible and can be extended in many ways. For example, the results of this paper have been generalized by Kang and Ramanan in [12] and [13] to include abandonments and also to establish ergodicity of many-server queues with abandonment. In addition, the characterization of the pre-limit obtained here is used to establish functional central limit theorems by Kaspi and Ramanan [14]. In the context of single-server queueing networks, there have been several recent works that have used measure-valued processes to study fluid limits (see [3, 6, 7] and references therein). In these papers, the measure-valued processes keep track of the residual service times or lead times of customers. In contrast, in the present paper we introduce a different measure-valued representation that

keeps track of the ages of customers in service. The latter representation offers several advantages such as yielding semimartingale representations that are more amenable to computation, and can therefore be more convenient in many contexts.

The outline of the paper is as follows. A precise mathematical description of the model, including the basic assumptions, is provided in Section 2. Section 3 introduces the fluid equations and contains a summary of the main results. Uniqueness of solutions to the fluid equations is established in Section 4, and the functional strong law of large numbers limit is proved in Section 5. Finally, the large-time or equilibrium behavior of the fluid limit is described in Section 6. In the remainder of this section, we introduce some common notation used in the paper.

1.2. *Notation and terminology.* The following notation will be used throughout the paper. $\mathbb{N}$ is the set of pbositive integers, $\mathbb{Z}_+$ is the set of nonnegative integers, $\mathbb{R}$ is set of real numbers and $\mathbb{R}_+$ the set of nonnegative real numbers. For $a, b \in \mathbb{R}$, $a \vee b$ denotes the maximum of a and b, $a \wedge b$ the minimum of a and b and the short-hand a^+ is used for $a \vee 0$. Given $A \subset \mathbb{R}$ and $a \in \mathbb{R}$, $A - a$ equals the set $\{x \in \mathbb{R} : x + a \in A\}$, and $\mathbb{1}_B$ denotes 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].

1.2.1. *Function spaces.* Given any Polish space S, $\mathcal{C}_b(S)$ and $\mathcal{C}_c(S)$ are, respectively, the space of bounded and continuous real-valued functions and the space of continuous real-valued functions with compact support defined on S. In this paper, S will typically be either a subset of $\mathbb{R}$ or of $\mathbb{R}^2_+$. Specifically, the space $\mathcal{C}_c([0, M) \times \mathbb{R}_+)$ will refer to the space of continuous functions on $\mathbb{R}^2$ that have compact support in $[0, M) \times \mathbb{R}_+$, restricted to the domain $[0, M) \times \mathbb{R}_+$, the space $\mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$ is defined to be the subset of functions φ in $\mathcal{C}_c([0, M) \times \mathbb{R}_+)$ for which the directional derivative

$$\varphi_x(x, s) + \varphi_s(x, s) \doteq \lim_{\Delta \downarrow 0} \frac{\varphi(x + \Delta, s + \Delta) - \varphi(x, s)}{\Delta}, \qquad (x, s) \in [0, M) \times \mathbb{R}_+,$$

exists and lies in $\mathcal{C}_c([0, M) \times \mathbb{R}_+)$, and the space $\mathcal{C}_b^{1,1}([0, M) \times \mathbb{R}_+)$ is the space of bounded functions on $[0, M) \times \mathbb{R}_+$ for which the directional derivative $\varphi_x + \varphi_s$ exists, and is also bounded and continuous. Similarly, the space $\mathcal{C}_c[0, M)$ is the space of continuous functions on $\mathbb{R}$ with compact support on $[0, M)$, restricted to the domain $[0, M)$. Moreover, $\mathcal{C}_c^1[0, M)$ is the subset of once continuously differentiable functions in $\mathcal{C}_c[0, M)$ and for $f \in \mathcal{C}_c^1[0, M)$, we let f' denote the derivative of f. In addition $\mathcal{C}_c^\infty(\mathbb{R}^2_+)$ is the space of infinitely differentiable functions on $\mathbb{R}^2_+$. Also, $\mathcal{B}V_0[0, \infty)$ is the space of càdlàg functions on $[0, \infty)$ with $f(0) = 0$ that have finite variation on every bounded interval in $\mathbb{R}_+$, and $\mathcal{I}_0[0, \infty)$ is the subspace of nondecreasing càdlàg functions with $f(0) = 0$. Let $\mathcal{L}^1[0, M)$ and $\mathcal{L}^1_{\text{loc}}[0, M)$ represent, respectively, the spaces of Lebesgue integrable and locally Lebesgue integrable functions on $[0, M)$. Recall that for $M \le \infty$, a function is said to be locally Lebesgue integrable on $[0, M)$ if and only if it satisfies $\int_{[0,m]} |f(x)|\, dx < \infty$

for all $m < M$. The constant functions $f \equiv 1$ and $f \equiv 0$ on $[0, M)$ will be represented by the symbols $\mathbf{1}$ and $\mathbf{0}$, respectively. With some abuse of notation, we will also use $\mathbf{1}$ and $\mathbf{0}$ for the constant functions on $[0, M) \times [0, \infty)$ that equal 1 and 0, respectively. The use will be clear from the context. In addition, we use $\tilde{\mathbf{1}}$ to denote the constant function on $[0, \infty)$ that is equal to 1. Furthermore, we use id to represent the identity function on $[0, \infty) : \mathrm{id}(t) = t$ for $t \in [0, \infty)$. Given any f defined on $[0, M)$, $M \le \infty$, we define $\|f\|_T \doteq \sup_{s\in[0,T)} |f(s)|$ for every $T \le M$. For a real-valued function φ on S, let $\|\varphi\|_\infty \doteq \sup_{x\in S} |\varphi(x)|$. Note that both $\|f\|_T$ and $\|f\|_\infty$ could possibly equal infinity. In addition, the support of a function φ is denoted by $\mathrm{supp}(\varphi)$.

Given a nondecreasing, right continuous function f with $f_* \doteq \sup_{s\in[0,\infty)} f(s)$, consider the following inverse functionals that take values in the extended reals:

$$\mathrm{inv}[f](t) \doteq \inf\{s \ge 0 : f(s) \ge t\}, \qquad t \in [0, f_*], \tag{1.1}$$

with the convention that $[0, f_*] = [0, \infty)$ if $f_* = \infty$, and if $f_* < \infty$ then $\mathrm{inv}[f](t) \doteq \infty$ for $t > f_*$. Likewise, let

$$f^{-1}(t) \doteq \sup\{s \ge 0 : f(s) \le t\}, \qquad t \in [0, f_*), \tag{1.2}$$

and $f^{-1}(t) \doteq \infty$ for $t \ge f_*$.

1.2.2. *Measure spaces.* The space of Radon measures on a Polish space S, endowed with the Borel σ-algebra, is denoted by $\mathcal{M}(S)$, while $\mathcal{M}_F(S)$, $\mathcal{M}_1(S)$ and $\mathcal{M}_{\le 1}(S)$ are, respectively, the subspaces of finite nonnegative, probability and sub-probability measures in $\mathcal{M}(S)$. Also, given $B < \infty$, $\mathcal{M}_{\le B}(S) \subset \mathcal{M}_F(S)$ denotes the space of measures μ in $\mathcal{M}_F(S)$ such that $\mu(S) \le B$. Recall that a Radon measure is one that assigns finite measure to every relatively compact subset of S. By identifying a Radon measure $\mu \in \mathcal{M}(S)$ with the mapping on $\mathcal{C}_c(S)$ defined by

$$\varphi \mapsto \int_S \varphi(x)\mu(dx),$$

one can equivalently define a Radon measure on S as a linear mapping from $\mathcal{C}_c(S)$ into $\mathbb{R}$ such that for every compact set $\mathcal{K} \subset S$, there exists $L_\mathcal{K} < \infty$ such that

$$\left| \int_S \varphi(x)\mu(dx) \right| \le L_\mathcal{K} \|\varphi\|_\infty \qquad \forall \varphi \in \mathcal{C}_c(S) \text{ with } \mathrm{supp}(\varphi) \subset \mathcal{K}. \tag{1.3}$$

The space $\mathcal{M}(S)$ is equipped with the vague topology, that is, a sequence of measures $\{\mu_n\}_{n\in\mathbb{N}}$ in $\mathcal{M}(S)$ is said to converge to μ in the vague topology (denoted $\mu_n \xrightarrow{v} \mu$) if and only if for every $\varphi \in \mathcal{C}_c(S)$,

$$\int_S \varphi(x)\mu_n(dx) \to \int_S \varphi(x)\mu(dx) \qquad \text{as } n \to \infty. \tag{1.4}$$

On $\mathcal{M}_F(S)$, we will also consider the weak topology, that is, a sequence $\{\mu_n\}$ in $\mathcal{M}_F(S)$ is said to converge weakly to μ (denoted $\mu_n \xrightarrow{w} \mu$) if and only if (1.4)

holds for every $\varphi \in \mathcal{C}_b(S)$. As is well known, $\mathcal{M}(S)$ and $\mathcal{M}_F(S)$, endowed with the vague and weak topologies, respectively, are Polish spaces. The symbol δ_x will be used to denote the measure with unit mass at the point x and, with some abuse of notation, we will use $\tilde{\mathbf{0}}$ to denote the identically zero Radon measure on S. When S is an interval, say $[0, M)$, for notational conciseness, we will often write $\mathcal{M}[0, M)$ instead of $\mathcal{M}([0, M))$.

As mentioned above, we will mostly be interested in the case when $S = [0, M)$ and $S = [0, M) \times [0, \infty)$, for some $M \in (0, \infty]$. To distinguish these cases, we will usually use f to denote generic functions on $[0, M)$ and φ to denote generic functions on $[0, M) \times [0, \infty)$. With some abuse of notation, given f on $[0, M)$, we will sometimes also treat it as a function on $[0, M) \times [0, \infty)$ that is constant in the second variable. For any Borel measurable function $f : [0, M) \to \mathbb{R}$ that is integrable with respect to $\xi \in \mathcal{M}[0, M)$, we often use the short-hand notation

$$\langle f, \xi \rangle \doteq \int_{[0,M)} f(x)\xi(dx).$$

Also, for ease of notation, given $\xi \in \mathcal{M}[0, M)$ and an interval $(a, b) \subset [0, M)$, we will use $\xi(a, b)$ and $\xi(a)$ to denote $\xi((a, b))$ and $\xi(\{a\})$, respectively.

1.2.3. *Measure-valued stochastic processes.* Given a Polish space $\mathcal{H}$, we let $\mathcal{D}_{\mathcal{H}}[0, T]$ and $\mathcal{D}_{\mathcal{H}}[0, \infty)$, respectively, denote the spaces of $\mathcal{H}$-valued, càdlàg functions on $[0, T]$ and $[0, \infty)$, both endowed with the usual Skorokhod J_1-topology [16]. Then $\mathcal{D}_{\mathcal{H}}[0, T]$ and $\mathcal{D}_{\mathcal{H}}[0, \infty)$ are also Polish spaces (see [16]). In this work, we will be interested in $\mathcal{H}$-valued stochastic processes, where $\mathcal{H} = \mathcal{M}_F[0, M)$ for some $M \leq \infty$. These are random elements that are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in $\mathcal{D}_{\mathcal{H}}[0, \infty)$, equipped with the Borel σ-algebra (generated by open sets under the Skorokhod J_1-topology). A sequence $\{X_n\}_{n\in\mathbb{N}}$ of càdlàg, $\mathcal{H}$-valued processes, with X_n defined on the probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$, is said to converge in distribution to a càdlàg $\mathcal{H}$-valued process X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ if, for every bounded, continuous functional $F : \mathcal{D}_{\mathcal{H}}[0, \infty) \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \mathbb{E}_n[F(X_n)] = \mathbb{E}[F(X)],$$

where $\mathbb{E}_n$ and $\mathbb{E}$ are the expectation operators with respect to the probability measures $\mathbb{P}_n$ and $\mathbb{P}$, respectively. Convergence in distribution of X_n to X will be denoted by $X_n \Rightarrow X$.

2. Model dynamics and basic assumptions. In Section 2.1 we describe our basic model and state our main assumptions, and in Section 2.2 we introduce some auxiliary processes that are useful for the study of the dynamics of the model.

2.1. *Description of the model.* Consider a system with N servers, where arriving customers are served in a nonidling, First-Come-First-Serve (FCFS) man-

ner, that is, a newly arriving customer immediately enters service if there are any idle servers, or, if all servers are busy, then the customer joins the back of the queue, and the customer at the head of the queue (if one is present) enters service as soon as a server becomes free. Our results are not sensitive to the exact mechanism used to assign an arriving customer to an idle server, as long as the nonidling condition is satisfied. Let $E^{(N)}$ denote the cumulative arrival process, with $E^{(N)}(t)$ representing the total number of customers that arrive into the system in the time interval $[0, t]$, and let the service requirements be drawn from an i.i.d. sequence $\{v_i, i = -N+1, -N+2, \ldots, 0, 1, \ldots\}$, with common cumulative distribution function G. For $i \in \mathbb{N}$, v_i represents the service requirement of the ith customer to enter service after time 0. Let $X^{(N)}(0)$ denote the number of customers in the system at time 0. Then $\{v_i, i = -(X^{(N)}(0) \wedge N) + 1, \ldots, 0\}$ represents the service requirements of customers already in service at time zero. When $E^{(N)}$ is a renewal process, this is simply a $GI/GI/N$ queueing system.

Consider the càdlàg, real-valued process $R_E^{(N)}$ defined by

$$R_E^{(N)}(s) \doteq \inf\{t > s : E^{(N)}(t) > E^{(N)}(s)\} - s, \tag{2.1}$$

which denotes the time from s until the next arrival. If $E^{(N)}$ is a renewal process, then $R_E^{(N)}$ is simply the forward recurrence time. The following mild assumptions will be imposed throughout, without explicit mention:

- $E^{(N)}$ is a nondecreasing, pure jump process with $E^{(N)}(0) = 0$ and for $t \in [0, \infty)$, $E^{(N)}(t) < \infty$ and $E^{(N)}(t) - E^{(N)}(t-) \in \{0, 1\}$;
- the cumulative arrival process $E^{(N)}$ is independent of the sequence of service requirements $\{v_j, j = -N+1, -N+2, \ldots\}$;
- the process $R_E^{(N)}$ is Markovian with respect to its own natural filtration; this holds, for example, when $E^{(N)}$ is a renewal process (see Proposition 1.5 of Section V of [1]) or an inhomogeneous Poisson process;
- G has density g;
- without loss of generality, we can (and will) assume that the mean service requirement is 1, that is

$$\int_{[0,\infty)} \big(1 - G(x)\big)\, dx = \int_{[0,\infty)} x g(x)\, dx = 1. \tag{2.2}$$

The sequence of processes $\{R_E^{(N)}, E^{(N)}, X^{(N)}(0), v_i, i = -N+1, \ldots, 0, 1, \ldots\}_{N \in \mathbb{N}}$ are all assumed to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that is large enough for the independence assumptions stated above to hold.

The first three assumptions stated above are very general, allowing for a large class of arrival processes. Note that the fourth assumption implies, in particular, that $G(0+) = 0$. The existence of a density is assumed for convenience, and is satisfied by a large class of distributions of interest in applications. The relaxation of this assumption would lead to a more complicated and somewhat less intuitive

representation for the fluid limit, and thus, for ease of exposition, we have restricted ourselves to this generality. Define the hazard rate

$$h(x) \doteq \frac{g(x)}{1 - G(x)}, \qquad x \in [0, M), \tag{2.3}$$

where

$$M \doteq \sup\{x \in [0, \infty) : G(x) < 1\}. \tag{2.4}$$

Note that in many interesting cases, $M = \infty$. Also, observe that h is always locally integrable on $[0, M)$ because for every $0 \le a \le b < M$,

$$\int_a^b h(x)\,dx = \ln\bigl(1 - G(a)\bigr) - \ln\bigl(1 - G(b)\bigr) < \infty.$$

However, the same calculation shows that h is not integrable on $[0, M)$. When additional assumptions on h are needed, they will be mentioned explicitly in the statements of the results.

The N-server model described above can be represented in many ways (see, e.g., representations for $GI/G/N$ queueing systems in [1], Chapter XII). For our purposes, we will find it convenient to encode the state of the system in the processes $(R_E^{(N)}, X^{(N)}, \nu^{(N)})$, where $R_E^{(N)}$ is the process defined in (2.1), $X^{(N)}(t)$ represents the number of customers in the system at time t (including those in service and those in the queue, waiting to enter service) and $\nu_t^{(N)}$ is the discrete nonnegative Borel measure on $[0, M)$ that has a unit mass at the age of each of the customers in service at time t. Here, the age $a_j^{(N)}$ of customer j is (for every realization) the piecewise linear function on $[0, \infty)$ that is defined to be 0 till the customer enters service, then increases linearly while the customer is in service (representing the amount of time elapsed since entering service) and is then constant (equal to the total service requirement) after the customer departs. Hence, the total number of customers in service at time t is given by $\langle \mathbf{1}, \nu_t^{(N)} \rangle = \nu_t^{(N)}[0, M)$, which is bounded above by N and so $\nu_t^{(N)} \in \mathcal{M}_{\le N}[0, M)$ for every $t \in [0, \infty)$. Our results will be independent of the particular rule used to assign customers to servers, but for technical purposes we will find it convenient to also introduce the additional "station process" sequence $\Upsilon^{(N)} \doteq (\Upsilon_j^{(N)}, j \in \{-(X^{(N)}(0) \wedge N) + 1, \ldots, 0\} \cup \mathbb{N})$, $N \in \mathbb{N}$, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each $t \in [0, \infty)$, if customer j has already entered service by time t, then $\Upsilon_j^{(N)}(t)$ is equal to the index $i \in \{1, \ldots, N\}$ of the station at which customer j receives/received service and $\Upsilon_j^{(N)}(t) \doteq 0$ otherwise. Finally, for $t \in [0, \infty)$, let $\tilde{\mathcal{F}}_t^{(N)}$ be the σ-algebra generated by $\{X^{(N)}(0), R_E^{(N)}(s), a_j^{(N)}(s), \Upsilon_j^{(N)}(s), j \in \{-(X^{(N)}(0) \wedge N) + 1, \ldots, 0\} \cup \mathbb{N}, s \in [0, t]\}$ and let $\{\mathcal{F}_t^{(N)}\}_{t \ge 0}$ denote the associated right continuous filtration, which is completed (with respect to $\mathbb{P}$) so that it satisfies the usual conditions. As discussed in the next section, it is not hard to see that $(E^{(N)}, X^{(N)}, \nu^{(N)})$ is $\{\mathcal{F}_t^{(N)}\}$-adapted. An explicit construction of these processes (in a more general setting

that allows the possibility of abandonments) can be found in Appendix A of Kang and Ramanan [12]. Moreover, it follows from Appendix B of [12] that $\{(R_E^{(N)}(t), X^{(N)}(t), \nu_t^{(N)}), \mathcal{F}_t^{(N)}, \mathbb{P}\}$ is a strong Markov process, although we do not use this property in this paper.

2.2. *Some auxiliary processes.* We now introduce the following auxiliary processes that will be useful for the study of the evolution of the system:

- the cumulative departure process $D^{(N)}$, where $D^{(N)}(0) = 0$ and for $t > 0$, $D^{(N)}(t)$ is the cumulative number of customers that have departed the system in the interval $(0, t]$;
- the process $K^{(N)}$, where $K^{(N)}(0) = 0$ and for $t > 0$, $K^{(N)}(t)$ represents the cumulative number of customers that have entered service in the interval $(0, t]$.

Simple mass balances show that

$$D^{(N)} \doteq X^{(N)}(0) - X^{(N)} + E^{(N)} \tag{2.5}$$

and

$$K^{(N)} \doteq \langle \mathbf{1}, \nu^{(N)} \rangle - \langle \mathbf{1}, \nu_0^{(N)} \rangle + D^{(N)}. \tag{2.6}$$

Due to the FCFS nature of the service and the absence of abandonments, observe that $K^{(N)}(t)$ is also the highest index of any customer that has entered service by time t, and so $K^{(N)}$ is $\{\mathcal{F}_t^{(N)}\}$-adapted.

For $N \in \mathbb{N}$ and each j, let $\alpha_j^{(N)} \doteq \text{inv}[K^{(N)}](j)$, where inv is defined as in (1.2). In other words, $\alpha_j^{(N)}$ is the time at which customer j enters service. The age process of each customer in service increases linearly, and so, given the service requirement of the customer, the evolution of the age process can be described explicitly in terms of the stopping times $\alpha_j^{(N)}$ as follows:

$$a_j^{(N)}(t) = \begin{cases} [t - \alpha_j^{(N)}] \vee 0, & \text{if } t - \alpha_j^{(N)} < v_j, \\ v_j, & \text{otherwise.} \end{cases} \tag{2.7}$$

For $t \geq 0$, the measure $\nu_t^{(N)}$ can be written in the form

$$\nu_t^{(N)} = \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{K^{(N)}(t)} \delta_{a_j^{(N)}(t)} \mathbb{1}_{\{a_j^{(N)}(t) < v_j\}}. \tag{2.8}$$

Recall that δ_x represents the Dirac mass at the point x. Now, at any time t, the age process of any customer has a right-derivative that is positive (and equal to one) if and only if the customer is in service, and has a left-derivative that is positive and a right-derivative that is zero if and only if it has just departed. Thus $D^{(N)}$ is

clearly $\{\mathcal{F}_t^{(N)}\}$-adapted and, since $\nu_t^{(N)}$ can be written explicitly purely in terms of the age process as

$$\nu_t^{(N)} = \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)}\rangle + 1}^{K^{(N)}(t)} \delta_{a_j^{(N)}(t)} \mathbb{1}_{\{d/dt a_j^{(N)}(t+) > 0\}}, \tag{2.9}$$

$\nu^{(N)}$ is also $\{\mathcal{F}_t^{(N)}\}$-adapted. Furthermore, since $\langle \mathbf{1}, \nu_t^{(N)}\rangle$ represents the number of customers in service at time t, the nonidling condition takes the form

$$N - \langle \mathbf{1}, \nu^{(N)}\rangle = \big[N - X^{(N)}\big]^+. \tag{2.10}$$

This shows that $\langle \mathbf{1}, \nu_t^{(N)}\rangle < N$ if and only if $X^{(N)}(t) < N$, which occurs if and only if the number in system is equal to the number in service, and so there is no queue. From the above discussion, it follows immediately that the processes $X^{(N)}$, $\nu^{(N)}$, $D^{(N)}$ and $K^{(N)}$ are all $\{\mathcal{F}_t^{(N)}\}$-adapted. For an explicit construction of these processes, in a more general context that allows the possibility of customer abandonment, see Appendix A of Kang and Ramanan [12].

REMARK 2.1. If $M < \infty$, then for every N, we will always assume that $\nu_0^{(N)}$ has support in $[0, M)$. From (2.8), this automatically implies that $\nu_t^{(N)}$ also has support in $[0, M)$ for every $t \in [0, \infty)$ and, moreover, that $\nu^{(N)} \in \mathcal{D}_{\mathcal{M}_{\le N}[0,M)}[0, \infty)$.

3. Main results. We now summarize our main results. First, in Section 3.1, we introduce the so-called fluid equations, which provide a continuous analog of the discrete model introduced in Section 2. In Section 3.2 we present our main results, which in particular show that, under the specified assumptions, the fluid equations uniquely characterize the strong law of large numbers limit of the multi-server system, as the number of servers goes to infinity. Lastly, in Section 3.3, we show how our results can be used to obtain fluid limits of various other functionals of interest. This, in particular, illustrates the usefulness of adopting a measure-valued representation for the state.

3.1. *Fluid equations.* Consider the following scaled versions of the basic processes describing the model. For $N \in \mathbb{N}$, the scaled state descriptor $(\overline{R}_E^{(N)}, \overline{X}^{(N)}, \overline{\nu}^{(N)})$ is given by

$$\overline{R}_E^{(N)}(t) \doteq R_E^{(N)}(t); \qquad \overline{X}^{(N)}(t) \doteq \frac{X^{(N)}(t)}{N}; \qquad \overline{\nu}_t^{(N)}(B) \doteq \frac{\nu_t^{(N)}(B)}{N} \tag{3.1}$$

for $t \in [0, \infty)$ and any Borel subset B of $[0, M)$, and observe that $\overline{\nu}_t^{(N)}$ is a sub-probability measure on $[0, M)$ for every $t \in [0, \infty)$. Analogously, define

$$\overline{E}^{(N)} \doteq \frac{E^{(N)}}{N}; \qquad \overline{D}^{(N)} \doteq \frac{D^{(N)}}{N}; \qquad \overline{K}^{(N)} \doteq \frac{K^{(N)}}{N}. \tag{3.2}$$

Recall that $\mathcal{I}_0[0,\infty)$ is the subset of nondecreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $f(0)=0$ and $M = \sup\{x \in [0,\infty) : G(x) < 1\}$, and define

$$\mathcal{S}_0 \doteq \{(f,x,\mu) \in \mathcal{I}_0[0,\infty) \times \mathbb{R}_+ \times \mathcal{M}_{\le 1}[0,M) : 1 - \langle \mathbf{1}, \mu\rangle = [1-x]^+\}. \tag{3.3}$$

$\mathcal{S}_0$ serves as the space of possible input data for the fluid equations. We make the following convergence assumptions on the primitives of the scaled sequence of systems.

ASSUMPTION 1 (Initial conditions). There exists an $\mathcal{S}_0$-valued random element $(\overline{E}, \overline{X}(0), \overline{\nu}_0)$ such that, as $N \to \infty$, the following limits hold:

(1) $\overline{E}^{(N)} \to \overline{E}$ in $\mathcal{D}_{\mathbb{R}_+}[0,\infty)$, $\mathbb{P}$-a.s., and $\mathbb{E}[\overline{E}^{(N)}(t)] \to \mathbb{E}[\overline{E}(t)] < \infty$ for every $t \in [0,\infty)$;
(2) $\overline{X}^{(N)}(0) \to \overline{X}(0)$ in $\mathbb{R}_+$, $\mathbb{P}$-a.s., and $\mathbb{E}[\overline{X}^{(N)}(0)] \to \mathbb{E}[\overline{X}(0)] < \infty$;
(3) $\overline{\nu}_0^{(N)} \to \overline{\nu}_0$ weakly in $\mathcal{M}_{\le 1}[0,M)$, $\mathbb{P}$-a.s.

REMARK 3.1. Note that conditions (1) and (2) of Assumption 1 imply that for every $t \in [0,\infty)$, $\limsup_N \mathbb{E}[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(t)] < \infty$.

REMARK 3.2. Using the Skorokhod representation theorem in the standard way, it can be shown that all the stochastic process convergence results in the paper continue to hold if, in Assumption 1, the almost sure limits are replaced by limits in the sense of weak convergence.

Our goal is to identify the limit in distribution of the quantities $(\overline{X}^{(N)}, \overline{\nu}^{(N)})$, as $N \to \infty$. In this section, we first introduce the so-called fluid equations and provide some intuition as to why the limit of any sequence $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)})\}$ should be expected to be a solution to these equations. In Section 5, we provide a rigorous proof of this fact. In what follows, h is the hazard rate function defined in (2.3).

DEFINITION 3.3 (Fluid equations). The càdlàg function $(\overline{X}, \overline{\nu})$ defined on $[0,\infty)$ and taking values in $[0,\infty) \times \mathcal{M}_{\le 1}[0,M)$ is said to solve the *fluid equations* associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$ if and only if for every $t \in [0,\infty)$,

$$\int_0^t \langle h, \overline{\nu}_s\rangle \, ds < \infty, \tag{3.4}$$

and the following relations are satisfied: for every $\varphi \in \mathcal{C}_c^{1,1}([0,M) \times \mathbb{R}_+)$,

$$\begin{aligned} \langle \varphi(\cdot,t), \overline{\nu}_t\rangle &= \langle \varphi(\cdot,0), \overline{\nu}_0\rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \overline{\nu}_s\rangle \, ds \\ &\quad - \int_0^t \langle h(\cdot)\varphi(\cdot,s), \overline{\nu}_s\rangle \, ds + \int_{[0,t]} \varphi(0,s) \, d\overline{K}(s); \end{aligned} \tag{3.5}$$

$$\overline{X}(t) = \overline{X}(0) + \overline{E}(t) - \int_0^t \langle h, \overline{\nu}_s\rangle \, ds \tag{3.6}$$

and

$$1 - \langle \mathbf{1}, \overline{\nu}_t \rangle = [1 - \overline{X}(t)]^+, \tag{3.7}$$

where

$$\overline{K}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \int_0^t \langle h, \overline{\nu}_s \rangle \, ds. \tag{3.8}$$

We now provide an intuitive explanation for the form of the fluid equations. Suppose $(\overline{X}, \overline{\nu})$ solves the fluid equations associated with some $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$. Then, roughly speaking, for $x \in \mathbb{R}_+$, $\overline{\nu}_s(dx)$ represents the amount of mass (or limiting fraction of customers) whose age lies in the range $[x, x + dx)$ at time s. Since h is the hazard rate, $h(x)$ represents the fraction of mass with age x that would depart from the system at any time. Thus, the quantity $\langle h, \overline{\nu}_s \rangle$, which is finite for almost every s by (3.4), represents the departure rate of mass from the fluid system at time s, and the process $\overline{D}$ given by

$$\overline{D}(t) \doteq \int_0^t \langle h, \overline{\nu}_s \rangle \, ds, \qquad t \in [0, \infty), \tag{3.9}$$

represents the cumulative amount of departures from the fluid system. Because $\overline{E}$ is the limiting cumulative arrival rate of mass into the fluid system in the interval $[0, t]$, a simple mass balance yields the relation (3.6) [which is the analogue of (2.5) describing the N-server system]. Likewise, (3.7) and (3.8) are the fluid versions of the nonidling condition (2.10) and the mass balance relation (2.6), respectively. It is clear from (3.6) that $\overline{X}$ is continuous (resp., absolutely continuous) if $\overline{E}$ is continuous (resp., absolutely continuous).

Next, note that the fluid equation (3.5) implies, in particular, that for $f \in \mathcal{C}_c^1[0, M)$,

$$\langle f, \overline{\nu}_t \rangle = \langle f, \overline{\nu}_0 \rangle + \int_0^t \langle f', \overline{\nu}_s \rangle \, ds - \int_0^t \langle fh, \overline{\nu}_s \rangle \, ds + f(0)\overline{K}(t). \tag{3.10}$$

The difference $\langle f, \overline{\nu}_t \rangle - \langle f, \overline{\nu}_0 \rangle$ is caused by three different phenomena—evolution of the mass in the system, departures and arrivals—which are represented by the second, third and fourth terms, respectively, on the right-hand side of (3.10). Specifically, the second term on the right-hand side represents the change in $\overline{\nu}$ due to the fact that the ages of all customers in service increase at a constant rate 1, the third term represents the change due to departures of customers that have completed service and the last term on the right-hand side of (3.10) accounts for new customers entering service. Here $\overline{K}(t)$ represents the cumulative amount of mass that has entered service in the fluid system, and is multiplied by $f(0)$ because, by definition, any customer entering service has age 0 at the time of entry.

To close the section, we state a simple property, which we will sometimes refer to as the "nonanticipative" property, of solutions to the fluid limit that will be used

in Section 6. For this, we require the following notation: for any $t \in [0, \infty)$,

$$\overline{E}^{[t]} \doteq \overline{E}(t+\cdot) - \overline{E}(t), \qquad \overline{K}^{[t]} \doteq \overline{K}(t+\cdot) - \overline{K}(t),$$
$$\overline{X}^{[t]} \doteq \overline{X}(t+\cdot), \qquad \overline{\nu}^{[t]} \doteq \overline{\nu}_{t+\cdot}.$$

LEMMA 3.4. *Suppose $(\overline{X}, \overline{\nu})$ is a solution to the fluid equations for a given initial condition $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$, and $\overline{K}$ is the associated process that satisfies (3.8). Then for any $t \in [0, \infty)$, $(\overline{X}^{[t]}, \overline{\nu}^{[t]})$ is a solution to the fluid equations associated with the initial condition $(\overline{E}^{[t]}, \overline{X}(t), \overline{\nu}_t) \in \mathcal{S}_0$, and $\overline{K}^{[t]}$ is the corresponding process that satisfies (3.8), with $\overline{\nu}$ replaced by $\overline{\nu}^{[t]}$.*

The proof of the lemma involves straightforward algebraic manipulations of the fluid equations, and is thus omitted.

3.2. *Summary of main results.* Our first result concerns uniqueness of solutions to the fluid equations, which is established at the end of Section 4.1.

THEOREM 3.5. *Given any $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$, there exists at most one solution $(\overline{X}, \overline{\nu})$ to the associated fluid equations (3.4)–(3.7). Also, if $\overline{\nu}$ satisfies (3.4) then $(\overline{X}, \overline{\nu})$ is a solution to the fluid equations (3.5)–(3.8) if and only if $(\overline{X}, \overline{\nu})$ satisfies (3.6) and, for every $f \in \mathcal{C}_b(\mathbb{R}_+)$,*

$$\begin{aligned} \int_{[0,M)} f(x)\overline{\nu}_t(dx) = &\int_{[0,M)} f(x+t)\frac{1-G(x+t)}{1-G(x)}\overline{\nu}_0(dx) \\ &+ \int_{[0,t]} f(t-s)\big(1-G(t-s)\big)\, d\overline{K}(s), \end{aligned} \tag{3.11}$$

where $\overline{K}$ is given by (3.8). Moreover, if $\overline{E}$ is absolutely continuous with derivative a.e. equal to $\overline{\lambda}$, then $\overline{K}$ is also absolutely continuous and its derivative $\overline{\kappa}$ satisfies for a.e. $t \in [0, \infty)$,

$$\overline{\kappa}(t) \doteq \begin{cases} \overline{\lambda}(t), & \text{if } \overline{X}(t) < 1, \\ \overline{\lambda}(t) \wedge \langle h, \overline{\nu}_t \rangle, & \text{if } \overline{X}(t) = 1, \\ \langle h, \overline{\nu}_t \rangle, & \text{if } \overline{X}(t) > 1. \end{cases} \tag{3.12}$$

Furthermore, if both $\overline{\nu}_0$ and $\overline{E}$ are absolutely continuous, then $\overline{\nu}_t$ is absolutely continuous for every $t \in [0, \infty)$.

REMARK 3.6. If $\overline{E}$ is absolutely continuous with respect to Lebesgue measure on $[0, \infty)$, then so is $\langle f, \overline{\nu} \rangle$, and the solutions to the fluid equation are continuous (in the time parameter), as the word fluid suggests.

It is also possible to consider the case when the residual service times of the customers already in the system is distributed according to another distribution $\tilde{G}$. Indeed, it can be shown that in this case, the relations (3.11) and (3.12) continue to hold, but with G in the first integral on the right-hand side of (3.11) replaced by $\tilde{G}$.

Our next main result shows that, under a mild additional condition on the hazard rate function h stated as Assumption 2 below, a solution to the fluid equations exists and is the functional law of large numbers limit of the N-server system, as $N \to \infty$.

ASSUMPTION 2. There exists $m_0 < M$ such that h is either bounded or lower semicontinuous on (m_0, M).

THEOREM 3.7. *Suppose the initial conditions* $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$ *satisfy Assumption 1. Then the sequence* $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)})\}$ *is relatively compact. If, in addition, Assumption 2 holds, then a unique solution* $(\overline{X}, \overline{\nu})$ *to the fluid equations associated with* $(\overline{E}, \overline{X}(0), \overline{\nu}_0)$ *exists and* $(\overline{X}^{(N)}, \overline{\nu}^{(N)})$ *converges weakly, as* $N \to \infty$, *to* $(\overline{X}, \overline{\nu})$.

The proof of Theorem 3.7 is given at the end of Section 5.4. The key steps in the proof involve showing tightness of the sequence $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)})\}$, which is carried out in Section 5.3, characterizing the limit points of the sequence as solutions to the fluid equations, which is done in Section 5.4, and invoking the uniqueness of solutions to the fluid equations stated above in Theorem 3.5.

REMARK 3.8. Define $\overline{\nu}_*$ to be the measure on $[0, M)$ that is absolutely continuous with respect to Lebesgue measure, and has density $1 - G(x)$: for any Borel set $A \subset [0, M)$,

$$\overline{\nu}_*(A) \doteq \int_A \big(1 - G(x)\big)\, dx. \tag{3.13}$$

It is easy to verify that, when $\overline{E}$ is absolutely continuous with derivative $\overline{\lambda}(\cdot)$ almost everywhere bounded below by 1, $\overline{X}(0) = c \geq 1$ and $\overline{\nu}_0 = \overline{\nu}_*$, the pair $(\overline{X}, \overline{\nu})$ defined by

$$\overline{X}(t) = c + \overline{E}(t) - t, \qquad \overline{\nu}_t = \overline{\nu}_*, \qquad t \in [0, \infty),$$

satisfy (3.6), (3.11) and (3.12) with $\overline{K}(t) = t$ and $\kappa = 1$ and $\overline{K} = \mathrm{id}$, where recall that id denotes the identity function $\mathrm{id}(t) = t, t \geq 0$. In particular, if $\overline{E} = \mathrm{id}$ then $(c\tilde{\mathbf{1}}, \overline{\nu}_*)$ constitutes an invariant solution for the fluid equations; that is, if $\overline{X}(0) = c$ and $\overline{\nu}_0 = \overline{\nu}_*$ then $\overline{X}(t) = c$ and $\overline{\nu}_t = \overline{\nu}_*$ for all $t \geq 0$.

In the time homogeneous setting (i.e., with constant fluid arrival rate) it is therefore natural to ask whether the component $\overline{\nu}$ of the unique solution to the fluid equations (when it exists) converges to $\overline{\nu}_*$ in the large-time limit. This is the subject

of our last result. Also, recall that a family of finite, nonnegative measures $\{\mu_t\}_{t\in\mathbb{R}_+}$ is said to converge weakly, as $t \to \infty$, monotonically up to a finite, nonnegative measure μ if and only if for every nonnegative, bounded, continuous function f, the sequence of real numbers $\langle f, \mu_t\rangle$ increases, as $t \to \infty$, to $\langle f, \mu\rangle$.

THEOREM 3.9. *Suppose Assumption 2 is satisfied. Given* $(\overline{\lambda}\,\mathrm{id}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$ *with* $\overline{\lambda} \in [0, 1]$, *let* $(\overline{X}, \overline{\nu})$ *be the unique solution to the associated fluid equations. Then the following two properties are satisfied*:

(1) *if* $\overline{X}(0) = 0$ *then, as* $t \to \infty$, $\overline{X}(t) = \langle \mathbf{1}, \overline{\nu}_t\rangle$ *converges monotonically up to* $\overline{\lambda}$ *and* $\overline{\nu}_t$ *converges weakly monotonically up to* $\overline{\lambda}\overline{\nu}_*$;
(2) *if the service distribution has a second moment, then given any initial condition* $(\mathrm{id}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$, *as* $t \to \infty$, $\overline{\nu}_t$ *converges weakly to* $\overline{\nu}_*$, *that is, for every* $f \in \mathcal{C}_b[0, \infty)$,

$$\lim_{t\to\infty} \langle f, \overline{\nu}_t\rangle = \langle f, \overline{\nu}_*\rangle = \int_{[0,\infty)} f(x)\big(1 - G(x)\big)\,dx. \tag{3.14}$$

The proof of Theorem 3.9 is presented in Section 6. For the case $\overline{\lambda} < 1$, property 1 of Theorem 3.9 was stated as Theorem 7.3 of [22] without proof.

REMARK 3.10. Our main theorems hold for the majority of distributions that arise in practice, including the exponential, lognormal, phase type, uniform, Weibull and Pareto distributions. It does not, however, cover the deterministic distribution.

3.3. *Fluid limits of other functionals.* In the last section, we identified the fluid limit of the scaled number of customers in system. In fact, the fluid limit contains a lot more information. For instance, as a direct consequence of the continuous mapping theorem, Theorem 3.7 also identifies the limit, as $N \to \infty$, of the scaled queue length process $\overline{Q}^{(N)} = Q^{(N)}/N$, which is the normalized number of customers waiting in queue (and not in service) at any time: we have

$$\overline{Q}^{(N)} \doteq \overline{Q}^{(N)}(0) + \overline{E}^{(N)} - \overline{K}^{(N)} \quad \Rightarrow \quad \overline{Q} \doteq \overline{Q}(0) + \overline{E} - \overline{K}.$$

Below, we identify the fluid limits of other functionals of interest.

3.3.1. *Waiting time.* The waiting time functional is of particular interest in the context of call centers, where service targets are often specified in terms of the proportion of calls that experience a wait of less than some given level (see, e.g., [2]).

Recall the definitions of $\mathrm{inv}[f]$ and f^{-1} given in (1.1) and (1.2), respectively. Assuming the system starts empty, the waiting time $w^{(N)}(j)$ of the jth customer

in the Nth system is the time elapsed between arrival into the system and entry into service. This functional can be written explicitly as

$$w^{(N)}(j) \doteq \operatorname{inv}\big[K^{(N)}\big](j) - \operatorname{inv}\big[E^{(N)}\big](j), \qquad j \in \mathbb{N}. \tag{3.15}$$

Also, consider the related process defined on $[0,\infty)$ by $\overline{w}^{(N)}(t) \doteq w^{(N)}(E^{(N)}(t))$ and note that for $t \in [0,\infty)$,

$$\overline{w}^{(N)}(t) = \operatorname{inv}\big[\overline{K}^{(N)}\big]\big(\overline{E}^{(N)}(t)\big) - \operatorname{inv}\big[\overline{E}^{(N)}\big]\big(\overline{E}^{(N)}(t)\big).$$

Finally, let $\overline{w}$ be the process given by

$$\overline{w}(t) \doteq \overline{K}^{-1}(\overline{E}(t)) - t, \qquad t \in [0,\infty). \tag{3.16}$$

We will say a function $f \in \mathcal{D}[0,\infty)$ is uniformly strictly increasing if it is absolutely continuous and there exists $\theta > 0$ such that $f'(t) \geq \theta$ for all $t \in [0,\infty)$. Note that for any such function, $f^{-1}(f(t)) = t$ and f^{-1} is continuous on $[0,\infty)$. We have the following fluid limit result for the waiting times in the system.

THEOREM 3.11. *Suppose the conditions of Theorem 3.7 hold, and $\overline{E}$ is uniformly strictly increasing. If, in addition, $\overline{K}$ is continuous and uniformly strictly increasing, then $\overline{w}^{(N)} \Rightarrow \overline{w}$ as $N \to \infty$.*

PROOF. By Assumption 1 and Theorem 3.7, it follows that $\overline{E}^{(N)} \Rightarrow \overline{E}$ and $\overline{K}^{(N)} \Rightarrow \overline{K}$. Using the Skorokhod representation theorem, we can assume that the convergence in both cases is almost sure. When combined with the fact that $\overline{E}$ and $\overline{K}$ are uniformly strictly increasing, Lemma 4.10 of [18] shows that $\operatorname{inv}[f^{(N)}] \to f^{-1}$ (almost surely, uniformly on compact sets) for $f = \overline{E}$ and $\overline{K}$. Now, fix $T < \infty$ and $\omega \in \Omega$ such that these limits hold and also fix some $\varepsilon > 0$. Moreover, let $N_0 = N_0(\omega) < \infty$ be such that for all $N \geq N_0$,

$$\sup_{s \in [0, \overline{E}^{(N)}(T)]} \big[\operatorname{inv}\big[f^{(N)}\big](s) - f^{-1}(s)\big] \leq \varepsilon$$

for $f = \overline{E}, \overline{K}$. Then we have

$$\begin{aligned}
&\sup_{t\in[0,T]} \big|\operatorname{inv}\big[\overline{K}^{(N)}\big]\big(\overline{E}^{(N)}(t)\big) - \overline{K}^{-1}(\overline{E}(t))\big| \\
&\quad\leq \sup_{t\in[0,T]} \big|\operatorname{inv}\big[\overline{K}^{(N)}\big]\big(\overline{E}^{(N)}(t)\big) - \overline{K}^{-1}\big(\overline{E}^{(N)}(t)\big)\big| \\
&\qquad + \sup_{t\in[0,T]} \big|\overline{K}^{-1}\big(\overline{E}^{(N)}(t)\big) - \overline{K}^{-1}(\overline{E}(t))\big| \\
&\quad\leq \varepsilon + \sup_{t\in[0,T]} \big|\overline{K}^{-1}\big(\overline{E}^{(N)}(t)\big) - \overline{K}^{-1}(\overline{E}(t))\big|.
\end{aligned}$$

The continuity of $\overline{K}^{-1}$ and the fact that a.s., $\overline{E}^{(N)} \to \overline{E}$ u.o.c., as $N \to \infty$, together, ensure that a.s. $|\overline{K}^{-1}(\overline{E}^{(N)}) - \overline{K}^{-1}(\overline{E})| \to 0$ u.o.c. as $N \to \infty$. So, we have

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \left|\mathrm{inv}\big[\overline{K}^{(N)}\big]\big(\overline{E}^{(N)}(t)\big) - \overline{K}^{-1}(\overline{E}(t))\right| \le \varepsilon.$$

Sending $\varepsilon \to 0$, we infer that $\mathrm{inv}[\overline{K}^{(N)}] \circ \overline{E}^{(N)} \to \overline{K}^{-1} \circ \overline{E}$ uniformly on $[0, T]$. An analogous argument shows that $\mathrm{inv}[\overline{E}^{(N)}] \circ \overline{E}^{(N)} \to \mathrm{id}$, where recall that $\mathrm{id}: t \mapsto t$ is the identity mapping on $[0, \infty)$. When combined with the definition of $\overline{w}$, the theorem follows. $\square$

3.3.2. *Workload process.* The workload (or unfinished work) process $V^{(N)}$ is defined to be the amount of work in the N-server system (including the work of customers waiting in queue and the residual service of customers in service)

$$V^{(N)}(t) = \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)}\rangle+1}^{K^{(N)}(t)} \big(v_j - a_j^{(N)}(t)\big)\mathbb{1}_{\{a_j^{(N)}(t)<v_j\}} + \sum_{j=K^{(N)}(t)+1}^{X^{(N)}(0)-\langle \mathbf{1}, \nu_0^{(N)}\rangle+E^{(N)}(t)} v_j.$$

Let the scaled workload process $\overline{V}^{(N)}$ be defined in the usual fashion. We briefly outline below how the results and techniques of this paper may be used to characterize the limit $\overline{V}$ of the sequence $\{\overline{V}^{(N)}\}$ of scaled workload processes. A rigorous proof is beyond the scope of this paper.

Let $\eta^{(N)}$ be the measure-valued process (analogous to $\nu^{(N)}$) that represents the residual service times (rather than the ages) of customers in service in the Nth system: for $t \in [0, \infty)$,

$$\eta_t^{(N)} \doteq \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)}\rangle+1}^{K^{(N)}(t)} \delta_{v_j - a_j^{(N)}(t)} \mathbb{1}_{\{a_j^{(N)}(t)<v_j\}}.$$

Also, let $\overline{\eta}^{(N)}$ denote the corresponding scaled quantity. Fluid equations can be derived for the limit $\overline{\eta}$ of the sequence $\{\overline{\eta}^{(N)}\}$ in a manner similar to those derived for $\overline{\nu}$ in this paper. Moreover, under mild assumptions, we believe it can be shown that, as $N \to \infty$, $\overline{\eta}^{(N)} \Rightarrow \overline{\eta}$, where for every $f \in C_c[0, M)$ and $t \in [0, \infty)$,

$$\langle f, \overline{\eta}_t\rangle \doteq \int_{[0,M)} \left(\int_0^\infty \frac{g(x+r)}{1-G(x)} f(r)\, dr\right) \overline{\nu}_t(dx). \tag{3.17}$$

A completely rigorous proof of this result is beyond the scope of this paper. However, below we provide a plausible argument to justify the above claim. Given the age x of any customer that was already in service at time 0, the probability that the residual service time of the customer at time t is greater than u is given by $(1 - G(x + t + u))/(1 - G(x))$. Thus the density of the residual service time distribution at time t for a customer that had age x at time 0 is $g(x+t+\cdot)/(1-G(x))$.

Likewise, the density of the residual service distribution at time t for a customer that entered the system at time $0 < s < t$ is $g(t - s + \cdot)$. Moreover, given the ages of all customers in service, the residual service times of customers in service are independent. Therefore, by a strong law of large numbers reasoning, one expects that the limiting residual service measure $\overline{\eta}$ can be written in terms of the limiting initial age measure $\overline{\nu}_0$ and limiting cumulative entry-into-service process $\overline{K}$ as follows: for $f \in C_c[0, M)$,

$$\langle f, \overline{\eta}_t \rangle = \int_{[0,M)} \left(\int_0^\infty \frac{g(x+t+r)}{1-G(x)} f(r)\, dr \right) \overline{\nu}_0(dx) + \int_{[0,t]} \left(\int_0^\infty g(t-s+r) f(r)\, dr \right) d\overline{K}(s).$$

The expression (3.17) can then be obtained by using the representation (4.3) to rewrite the right-hand side above as an integral with respect to $\overline{\nu}_t$.

From the definition of $\eta_t^{(N)}$, the workload process admits the alternative representation

$$V^{(N)}(t) = \int_{[0,M)} x \eta_t^{(N)}(dx) + \sum_{j=K^{(N)}(t)+1}^{E^{(N)}(t)+X^{(N)}(0)-\langle \mathbf{1}, \nu_0^{(N)} \rangle} v_j.$$

Due to (2.5), (2.6) and (2.10), it follows that the number of terms in the second sum equals $[X^{(N)}(t) - N]^+$. When combined with the fact that the service times $\{v_j\}$ are i.i.d. with mean 1 and the convergence of $\overline{\eta}^{(N)}$ to $\overline{\eta}$ it is natural to conjecture (under suitable assumptions that justify the substitution of linear test functions f) the convergence $\overline{V}^{(N)} \Rightarrow \overline{V}$ as $N \to \infty$, where

$$\overline{V}(t) \doteq \int_{[0,M)} \left(\int_0^\infty \frac{r g(x+r)}{1-G(x)}\, dr \right) \overline{\nu}_t(dx) + \left(\overline{X}(t) - 1\right)^+. \tag{3.18}$$

It is worthwhile to note that, when $\overline{\nu}_0$ equals the invariant measure $\overline{\nu}_*$ defined in (3.13), then $\overline{\nu}_t = \overline{\nu}_*$ for all $t \in [0, \infty)$ and $\overline{V}(t) < \infty$ if and only if G has a finite second moment.

4. Uniqueness of solutions to the fluid equation. In this section we show that there is at most one solution to the fluid equation for any given initial condition. In fact, we will establish two stronger properties of the fluid equation, both of which imply uniqueness. The first is continuity of the mapping that takes $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$ to a corresponding solution $(\overline{X}, \overline{\nu})$ of the fluid equation, which is established in Section 4.1. The second is a maximality property that is established in Section 4.2. The proofs of both continuity and maximality rely on identifying the solution to a certain integral equation, which is carried out in Section 4.3. Existence of solutions to the fluid equation will follow from results established in Section 5 (see, in particular, Theorem 5.15).

4.1. *Continuity of the fluid equation map.* We begin by analyzing the integral equation (4.2) below, which is the fluid equation (3.5), but with $\overline{K}$ replaced by an arbitrary, bounded variation càdlàg function Z in $\mathcal{B}V_0[0,\infty)$, and with $\overline{\nu}_0$ replaced by an arbitrary Radon measure υ_0 in $\mathcal{M}[0,M)$. Specifically, in Theorem 4.1 we provide an explicit formula for the solution $\overline{\nu}$ to the integral equation in terms of υ_0 and Z. Roughly speaking, when $\upsilon_0 = \overline{\nu}_0$ and Z is any càdlàg, nondecreasing process this formula characterizes the evolution of the fluid age process $\overline{\nu}$ that would result when the cumulative fluid arrivals into service is Z. On substituting $Z = \overline{K}$, this yields a relation that must be satisfied by any pair of processes $\overline{\nu}$ and $\overline{K}$ that satisfy the fluid equations for the initial condition $\overline{\nu}_0$. This relation, along with the nonidling condition, is then used to establish continuity of the fluid solution map in Theorem 4.6.

Recall that $M \in (0,\infty]$ is the right-end of the support of the hazard rate function h, and that h is always locally Lebsegue integrable on $[0,M)$.

THEOREM 4.1. *Suppose* $\{\overline{\nu}_s\}_{s\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ *has the property that for every* $m \in [0,M)$ *and* $T \in [0,\infty)$, *there exists* $C(m,T) < \infty$ *such that*

$$\left| \int_0^\infty \langle \varphi(\cdot,s) h(\cdot), \overline{\nu}_s \rangle \, ds \right| \leq C(m,T) \|\varphi\|_\infty \tag{4.1}$$

for every $\varphi \in \mathcal{C}_c([0,M)\times[0,\infty))$ *with* $\operatorname{supp}(\varphi) \subseteq [0,m]\times[0,T]$. *Then, given any* $\upsilon_0 \in \mathcal{M}[0,M)$ *and* $Z \in \mathcal{B}V_0[0,\infty)$, $\{\overline{\nu}_t\}_{t\geq 0}$ *satisfies the integral equation*

$$\begin{aligned} \langle \varphi(\cdot,t), \overline{\nu}_t \rangle = {} & \langle \varphi(\cdot,0), \upsilon_0 \rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \overline{\nu}_s \rangle \, ds \\ & - \int_0^t \langle h(\cdot)\varphi(\cdot,s), \overline{\nu}_s \rangle \, ds + \int_{[0,t]} \varphi(0,s) \, dZ(s) \end{aligned} \tag{4.2}$$

for every $\varphi \in \mathcal{C}_c^{1,1}([0,M)\times\mathbb{R}_+)$ *and* $t \in [0,\infty)$, *if and only if* $\{\overline{\nu}_s\}_{s\geq 0}$ *satisfies*

$$\begin{aligned} \int_{[0,M)} f(x) \overline{\nu}_t(dx) = {} & \int_{[0,M)} f(x+t) \frac{1-G(x+t)}{1-G(x)} \upsilon_0(dx) \\ & + \int_{[0,t]} f(t-s)\big(1-G(t-s)\big) \, dZ(s) \end{aligned} \tag{4.3}$$

for every $f \in \mathcal{C}_c(\mathbb{R}_+)$ *and* $t \in [0,\infty)$. *Moreover, if* $\upsilon_0 \in \mathcal{M}_F[0,M)$, *then* (4.3) *holds for every* $f \in \mathcal{C}_b(\mathbb{R}_+)$.

REMARK 4.2. We shall refer to the integral equation (4.2) as the *age equation* (corresponding to υ_0 and Z). Note that (4.1) is implied by condition (3.4) of the fluid equations and, as remarked earlier, the age equation is simply the fluid equation (3.5), with υ_0 and Z in place of $\overline{\nu}_0$ and $\overline{K}$, respectively. Furthermore, note that equation (4.3) only depends on the values of f in $[0,M)$ since $f(u)(1-G(u)) = 0$ for all $u \geq M$, and (4.3) completely characterizes the deterministic measure-valued process $\overline{\nu}$.

REMARK 4.3. The last integral in (4.3) is, as usual, to be interpreted as a Riemann–Stieltjes integral. A straightforward integration-by-parts shows that for every $f \in \mathcal{C}_b^1(\mathbb{R}_+)$ and $t \in [0,\infty)$, this integral also admits the alternative representation

$$\begin{aligned}&\int_{[0,t]} f(t-s)\big(1-G(t-s)\big)\,dZ(s)\\ &\quad= f(0)Z(t)+\int_{[0,t]} f'(t-s)\big(1-G(t-s)\big)Z(s)\,ds\\ &\qquad-\int_{[0,t]} f(t-s)g(t-s)Z(s)\,ds.\end{aligned}\tag{4.4}$$

The proof of Theorem 4.1 involves PDE techniques and is relegated to Section 4.3. As a simple corollary of Theorem 4.1, we have the following result.

COROLLARY 4.4. *Let $(\overline{X},\overline{\nu})$ be a solution to the fluid equations associated with $(\overline{E},\overline{X}(0),\overline{\nu}_0)\in\mathcal{S}_0$. Then, for every $t\in[0,\infty)$, the function $\overline{K}$ defined by* (3.8) *satisfies the renewal equation*

$$\begin{aligned}\overline{K}(t)&=\langle\mathbf{1},\overline{\nu}_t\rangle-\langle\mathbf{1},\overline{\nu}_0\rangle+\int_{[0,M)}\frac{G(x+t)-G(x)}{1-G(x)}\overline{\nu}_0(dx)\\ &\quad+\int_0^t g(t-s)\overline{K}(s)\,ds,\end{aligned}\tag{4.5}$$

and admits the representation

$$\begin{aligned}\overline{K}(t)&=\int_{[0,t]}\big(\langle\mathbf{1},\overline{\nu}_{t-s}\rangle-\langle\mathbf{1},\overline{\nu}_0\rangle\big)\,dU(s)\\ &\quad+\int_{[0,t]}\left(\int_{[0,M)}\frac{G(x+t-s)-G(x)}{1-G(x)}\overline{\nu}_0(dx)\right)dU(s),\end{aligned}\tag{4.6}$$

where dU is the renewal measure associated with the distribution G.

PROOF. We first claim that if $(\overline{X},\overline{\nu})$ solve the fluid equations associated with $(\overline{E},\overline{X}(0),\overline{\nu}_0)$, then $\overline{K}$ defined by (3.8) must necessarily be nondecreasing (as one would expect from the interpretation of $\overline{K}$ as the limiting fraction of cumulative entries into service). In order to justify the claim, fix $0\le s\le t$. If $\overline{X}(t)>1$ then the nonidling condition (3.7) implies that $\langle\mathbf{1},\overline{\nu}_t\rangle=1$ which, when substituted into (3.8), shows that

$$\overline{K}(t)-\overline{K}(s)=1-\langle\mathbf{1},\overline{\nu}_s\rangle+\int_s^t\langle h,\overline{\nu}_u\rangle\,du\ge\int_s^t\langle h,\overline{\nu}_u\rangle\,du\ge 0.$$

On the other hand, if $\overline{X}(t)\le 1$, then the nonidling condition (3.7) shows that $\langle\mathbf{1},\overline{\nu}_t\rangle=\overline{X}(t)$ and $\langle\mathbf{1},\overline{\nu}_s\rangle\le\overline{X}(s)$. Hence, (3.8), (3.6) and the fact that $\overline{E}$ is nondecreasing, show that

$$\overline{K}(t)-\overline{K}(s)=\langle\mathbf{1},\overline{\nu}_t\rangle-\overline{X}(t)-\langle\mathbf{1},\overline{\nu}_s\rangle+\overline{X}(s)+\overline{E}(t)-\overline{E}(s)\ge 0,$$

which proves the claim.

In addition, by assumption, $\overline{K}$ and $\overline{\nu}$ satisfy (3.4) and (3.5). In other words, (4.1) is satisfied and (4.2) holds with $\upsilon_0 = \overline{\nu}_0$ and $Z = \overline{K}$. Therefore, by Theorem 4.1, (4.3) also holds with $\upsilon_0 = \overline{\nu}_0$ and $Z = \overline{K}$, and substituting $f = \mathbf{1}$, we obtain

$$\begin{aligned}
&\int_{[0,t]} \bigl(1 - G(t-s)\bigr)\, d\overline{K}(s) \\
&\quad = \langle \mathbf{1}, \overline{\nu}_t \rangle - \int_{[0,M)} \frac{1 - G(x+t)}{1 - G(x)} \overline{\nu}_0(dx) \\
&\quad = \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \int_{[0,M)} \frac{G(x+t) - G(x)}{1 - G(x)} \overline{\nu}_0(dx).
\end{aligned}$$

On the other hand, equation (4.4), with $Z = \overline{K}$ and $f = \mathbf{1}$, shows that

$$\int_{[0,t]} \bigl(1 - G(t-s)\bigr)\, d\overline{K}(s) = \overline{K}(t) - \int_0^t g(t-s)\overline{K}(s)\, ds.$$

Equating the right-hand sides of the last two displays, we obtain (4.5). Finally, since $\overline{K}$ is bounded on finite intervals, and the sum of the first two terms on the right-hand side of (4.5) is uniformly bounded by two, representation (4.6) is a direct result of the renewal theorem (see, e.g., Theorem 2.4(ii) of Section V in [1]). □

As an immediate consequence of Theorem 4.1 and Corollary 4.4, we obtain the following simple bound. Given a Radon measure $\mu \in \mathcal{M}[0, M)$, let $|\mu|_{\mathrm{TV}}$ represent the total variation of μ on $[0, M)$.

LEMMA 4.5. *For $i = 1, 2$, suppose $\upsilon_0^i \in \mathcal{M}[0, M)$ and $Z^i \in \mathcal{B}V_0[0, \infty)$ are given, and suppose (4.1) and (4.2) are satisfied with $\overline{\nu}$, υ_0 and Z replaced by $\overline{\nu}^i$, υ_0^i and Z^i, respectively. Then for every $T < \infty$ and $f \in \mathcal{C}_b^1(\mathbb{R}_+)$,*

$$\|\langle f, \overline{\nu}_s^2 \rangle - \langle f, \overline{\nu}_s^1 \rangle\|_T \le \|f\|_M |\Delta\upsilon_0|_{\mathrm{TV}} + (2\|f\|_T + \|f'\|_T)\|\Delta Z\|_T, \tag{4.7}$$

where $\Delta Z \doteq Z^2 - Z^1$ and $\Delta\upsilon_0 \doteq \upsilon_0^{(2)} - \upsilon_0^{(1)}$.

PROOF. By Theorem 4.1 and Remark 4.3, for $i = 1, 2$, relations (4.3) and (4.4) are satisfied with $\overline{\nu}$, υ_0 and Z replaced by $\overline{\nu}^i$, υ_0^i and Z^i, respectively. Together, these relations imply that for $f \in \mathcal{C}_b^1(\mathbb{R}_+)$ and $t \in [0, \infty)$,

$$\begin{aligned}
\langle f, \overline{\nu}_t^2 \rangle - \langle f, \overline{\nu}_t^1 \rangle &= \int_{[0,M)} f(x+t) \frac{1 - G(x+t)}{1 - G(x)} \Delta\upsilon_0(dx) + f(0)\Delta Z(t) \\
&\quad + \int_0^t f'(t-s)\bigl(1 - G(t-s)\bigr)\Delta Z(s)\, ds \\
&\quad - \int_0^t f(t-s)g(t-s)\Delta Z(s)\, ds.
\end{aligned}$$

Since $1 - G(u) = 0$ for $u \geq M$, this implies that for $f \in \mathcal{C}_b^1(\mathbb{R}_+)$ and for every $t \in [0, T]$,

$$|\langle f, \overline{\nu}_t^2\rangle - \langle f, \overline{\nu}_t^1\rangle| \leq \int_{[0,M-t)} |f(x+t)||\Delta \upsilon_0|(dx) + \big(|f(0)| + \|f\|_{t\wedge M} + \|f'\|_{t\wedge M}\big)\|\Delta Z\|_T$$

from which (4.7) follows. □

We now state the main result of this section. Below, ΔH denotes $H^{(2)} - H^{(1)}$ for $H = \overline{K}, \overline{D}, \overline{E}, \overline{X}$ and $\overline{\nu}$.

THEOREM 4.6 (Continuity of solution map). *For $i = 1, 2$, let $(\overline{X}^i, \overline{\nu}^{(i)})$ be a solution to the fluid equations associated with $(\overline{E}^i, \overline{X}^i(0), \overline{\nu}_0^{(i)}) \in \mathcal{S}_0$ and let $\overline{K}^i$ and $\overline{D}^i$ be defined as in (3.8) and (3.9), respectively, with $\overline{\nu}$ replaced by $\overline{\nu}^i$. If $\overline{\nu}_0^1 = \overline{\nu}_0^2$ then for every $T < \infty$,*

$$\Big[\sup_{t\in[0,T]} \Delta\overline{K}(t)\Big] \vee \Big[\sup_{t\in[0,T]} \Delta\overline{D}(t)\Big] \leq \Big[|\Delta\overline{X}(0)| + \sup_{t\in[0,T]} \Delta\overline{E}(t)\Big] \vee 0 \tag{4.8}$$

and, hence,

$$\|\Delta\overline{K}\|_T \vee \|\Delta\overline{D}\|_T \leq |\Delta\overline{X}(0)| + \|\Delta\overline{E}\|_T. \tag{4.9}$$

Moreover, for every $T < \infty$ and $f \in \mathcal{C}_b^1(\mathbb{R}_+)$,

$$\|\langle f, \overline{\nu}_s^2\rangle - \langle f, \overline{\nu}_s^1\rangle\|_T \leq (2\|f\|_T + \|f'\|_T)\big(\Delta\overline{X}(0) + \|\Delta\overline{E}\|_T\big). \tag{4.10}$$

PROOF. Fix $T < \infty$ and define $\varepsilon \geq 0$ by

$$\varepsilon \doteq \Big[\Delta\overline{X}(0) + \sup_{t\in[0,T]} \Delta\overline{E}(t)\Big] \vee 0. \tag{4.11}$$

For $\delta > 0$, let

$$\tau_\delta \doteq \inf\{t \geq 0 : \Delta\overline{K}(t) \geq \varepsilon + \delta\}.$$

We shall prove by contradiction that $\tau_\delta > T$ a.s., from which the continuity property will follow. If $\tau_\delta = \infty$ for all $\delta > 0$, then $\sup_{t\in[0,\infty)} \Delta\overline{K}(t) \leq \varepsilon$, and the result follows. Therefore, we can assume without loss of generality that there exists $\delta > 0$ such that $\tau_\delta < \infty$, and, for simplicity of notation, denote τ_δ simply by τ. The right-continuity of $\overline{K}^1$ and $\overline{K}^2$ imply that

$$\Delta\overline{K}(\tau) \geq \varepsilon + \delta. \tag{4.12}$$

We now show that $\tau > T$. Indeed, suppose $\tau \in [0, T]$ and consider the following two cases:

Case 1. $\overline{X}^1(\tau) < 1$. In this case, the nonidling condition (3.7) implies that

$$\overline{X}^1(\tau) - \langle \mathbf{1}, \overline{\nu}_\tau^1\rangle = 0 \leq \overline{X}^2(\tau) - \langle \mathbf{1}, \overline{\nu}_\tau^2\rangle.$$

Together with relations (3.8), (3.6), (4.11) and the fact that $\Delta\overline{\nu}_0 \equiv 0$, this implies that

$$\Delta\overline{K}(\tau) = \Delta\overline{E}(\tau) + \Delta\overline{X}(0) - \langle \mathbf{1}, \Delta\overline{\nu}_0\rangle - \Delta\overline{X}(\tau) + \langle \mathbf{1}, \Delta\overline{\nu}_\tau\rangle \le \varepsilon,$$

which contradicts (4.12).

Case 2. $\overline{X}^1(\tau) \ge 1$. In this case, due to the nonidling condition (3.7), we have $\langle \mathbf{1}, \overline{\nu}^1_\tau\rangle = 1 \ge \langle \mathbf{1}, \overline{\nu}^2_\tau\rangle$. Along with Corollary 4.4 and the fact that $\Delta\overline{\nu}_0 = 0$, this implies that

$$\begin{aligned}\Delta\overline{K}(\tau) &= \langle \mathbf{1}, \Delta\overline{\nu}_\tau\rangle - \langle \mathbf{1}, \Delta\overline{\nu}_0\rangle + \int_{[0,M)} \frac{G(x+\tau) - G(x)}{1 - G(x)} \Delta\overline{\nu}_0(dx) \\ &\quad + \int_0^\tau g(\tau - s)\Delta\overline{K}(s)\,ds \\ &= \langle \mathbf{1}, \Delta\overline{\nu}_\tau\rangle + \int_0^\tau g(\tau - s)\Delta\overline{K}(s)\,ds \\ &\le \int_0^\tau g(\tau - s)\Delta\overline{K}(s)\,ds.\end{aligned}$$

We now assert that the right-hand side is *strictly* less than $\varepsilon + \delta$, which contradicts (4.12). To see why the assertion holds, note that if $g(s) = 0$ for a.e. $s \in [0, \tau]$, then the right-hand side of the last inequality equals zero, which is trivially strictly less than $\varepsilon + \delta$. On the other hand, if $g(s) > 0$ for a set of positive Lebesgue measure in $[0, \tau]$, then the fact that $\Delta\overline{K}(s) < \varepsilon + \delta$ for all $s \in [0, \tau)$ shows once again that

$$\Delta\overline{K}(\tau) \le \int_0^\tau g(\tau - s)\Delta\overline{K}(s)\,ds < (\varepsilon + \delta)G(\tau) \le (\varepsilon + \delta).$$

Thus, in both cases 1 and 2, we arrive at a contradiction. Hence, it must be that $\tau > T$, which means that $\Delta\overline{K}(t) < \varepsilon + \delta$ for every $t \in [0, T]$. Sending $\delta \downarrow 0$, we conclude that $\Delta\overline{K}(t) \le \varepsilon$ for $t \in [0, T]$, as desired. In turn, using the relations (3.8), (3.9) and Corollary 4.4, along with the identity $\Delta\overline{\nu}_0 \equiv 0$ and the nonnegativity of g, we obtain for every $t \in [0, T]$,

$$\Delta\overline{D}(t) = \Delta\overline{K}(t) - \langle \mathbf{1}, \Delta\overline{\nu}_t\rangle = \int_0^t g(t - s)\Delta\overline{K}(s)\,ds \le \varepsilon G(t) \le \varepsilon.$$

This completes the proof of (4.8), and relation (4.9) follows by symmetry. Finally, since for $i = 1, 2$, $\overline{\nu}^i$ and $\overline{K}^i$ satisfy the fluid equations (by assumption), inequality (4.10) is a direct consequence of Lemma 4.5 and inequality (4.9). $\square$

PROOF OF THEOREM 3.5. Let $(\overline{X}^1, \overline{\nu}^1)$ and $(\overline{X}^2, \overline{\nu}^2)$ be two solutions to the fluid equations corresponding to $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$. Fix $r \in [0, M)$ and choose a sequence of functions $f_n \in \mathcal{C}^1_b(\mathbb{R}_+)$, $n \in \mathbb{N}$, such that $f_n \uparrow \mathbb{1}_{[0,r)}$ pointwise as $n \to \infty$. Then for every $t \in [0, \infty)$ and $n \in \mathbb{N}$, $\langle f_n, \overline{\nu}^1_t\rangle = \langle f_n, \overline{\nu}^2_t\rangle$ due to (4.10) and the fact that $\overline{E}^1 = \overline{E}^2$ and $\overline{X}^1 = \overline{X}^2$. Sending $n \to \infty$ and invoking the monotone

convergence theorem, we conclude that $\overline{\nu}_t^1[0,r) = \overline{\nu}_t^2[0,r)$. Since r and t are arbitrary, it follows that $\overline{\nu}^1 = \overline{\nu}^2$ and hence, by (3.6), that $\overline{X}^1 = \overline{X}^2$. This shows that there is at most one solution to the fluid equations. The second assertion follows immediately from Theorem 4.1 and Remark 4.2.

Now, suppose $\overline{E}$ is absolutely continuous with derivative $\overline{\lambda}$. Then (3.6) immediately shows that $\overline{X}$ is also absolutely continuous. In turn, using (3.7), (3.8) and the fact that $|[1-a]^+ - [1-b]^+| \le |a-b|$, it is easy to see that $\overline{K}$ is also absolutely continuous. Fix t such that both the derivative $\overline{\lambda}$ of $\overline{E}$ and the derivative $\overline{\kappa}$ of $\overline{K}$ exist at t. If $\overline{X}(t) < 1$, then the nonidling condition (3.7) and the continuity of $\overline{X}$ show that $\langle \mathbf{1}, \overline{\nu}_s \rangle = \overline{X}(s)$ for all s in a neighborhood of t. When combined with (3.6) and (3.8), this shows that $\overline{\kappa}(t) = \overline{\lambda}(t)$. On the other hand, if $\overline{X}(t) > 1$, then (3.7) and the continuity of $\overline{X}$ show that $\langle \mathbf{1}, \overline{\nu}_s \rangle = 1$ for s in a neighborhood of t. When substituted into (3.8) this shows that $\overline{\kappa}(t) = \langle h, \overline{\nu}_t \rangle$. Finally, since $\overline{X}$ and $\langle \mathbf{1}, \overline{\nu} \rangle$ are absolutely continuous, $d\overline{X}(t)/dt = d\langle \mathbf{1}, \overline{\nu}_t \rangle/dt = 0$ for a.e. t on which $\overline{X}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle = 1$ (see, e.g., Theorem A.6.3 of [4]). Together with (3.6) and (3.8), this implies that for a.e. $t \in [0,\infty)$ such that $\overline{X}(t) = 1$ [and $\overline{\lambda}(t)$ and $\kappa(t)$ are well defined], we have $\overline{\kappa}(t) = \overline{\lambda}(t) = \langle h, \overline{\nu}_t \rangle = \overline{\lambda}(t) \wedge \langle h, \overline{\nu}_t \rangle$. This proves (3.12). Finally, because $\overline{K}$ is absolutely continuous, if $\overline{\nu}_0$ is also absolutely continuous then the representation (3.11) immediately guarantees that $\overline{\nu}_s$ is absolutely continuous for every $s \in [0,\infty)$. □

4.2. *A maximality property of the fluid solution.* In this section we establish a result of independent interest. This result is not used in the rest of the paper, and can thus be safely skipped without loss of continuity. Specifically, we show that the nonidling property (3.7) implies a certain maximality property for solutions to the fluid equations. In particular, this result provides an alternative proof of uniqueness of solutions to the fluid limit that is different from the one using continuity of the solution map given in the last section.

Let $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$. Suppose that $(\overline{X}, \overline{\nu})$ solve the corresponding fluid equations (3.4)–(3.7), and let $\overline{K}$ and $\overline{D}$ be the associated processes, as defined in (3.8) and (3.9), respectively. Also, let $(\overline{X}^\diamond, \overline{\nu}^\diamond)$ be any process taking values in $\mathbb{R}_+ \times \mathcal{M}_{\le 1}[0,M)$ that satisfy the fluid equations, (3.4)–(3.6), and the relation

$$\langle \mathbf{1}, \overline{\nu}_t^\diamond \rangle \le \overline{X}^\diamond(t), \qquad t \in [0,\infty). \tag{4.13}$$

Here, $\overline{X}^\diamond$ and $\overline{\nu}^\diamond$, respectively, represent the total number of (fluid) customers in system and the distribution of ages of (fluid) customers in service under any given feasible assignment of customers to servers that does not necessarily satisfy the nonidling condition (3.7). Let $\overline{K}^\diamond$ and $\overline{D}^\diamond$, respectively, be the corresponding processes representing the cumulative entry into service and cumulative departures from the system, as defined by the right-hand sides of (3.8) and (3.9), respectively, but with $\overline{\nu}$ replaced by $\overline{\nu}^\diamond$. Then we have the following intuitive result that shows that the nonidling condition (3.7) ensures that the cumulative entry into service and cumulative departures from the system are maximized.

LEMMA 4.7. *For every* $t \in [0, \infty)$, $\overline{K}(t) \geq \overline{K}^{\diamond}(t)$ *and* $\overline{D}(t) \geq \overline{D}^{\diamond}(t)$.

PROOF. We shall argue by contradiction to prove the lemma. Fix $\varepsilon > 0$ and let

$$T = \inf\{t : \overline{K}^{\diamond}(t) \geq \overline{K}(t) + \varepsilon\}.$$

Suppose $T < \infty$. Then we consider the following two mutually exhaustive cases:

Case 1. $\overline{X}(T) < 1$. In this case, (3.7) implies that $\overline{X}(T) = \langle \mathbf{1}, \overline{\nu}_T \rangle$ which, along with (3.6) and (3.8), shows that

$$\overline{K}(T) = \overline{X}(0) - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \overline{E}(T).$$

On the other hand, (3.6), (3.8) and (4.13), when combined, show that for every $t \in [0, \infty)$,

$$\overline{K}^{\diamond}(t) = \langle \mathbf{1}, \overline{\nu}_t^{\diamond} \rangle - \overline{X}^{\diamond}(t) + \overline{X}(0) - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \overline{E}(t) \leq \overline{X}(0) - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \overline{E}(t).$$

The last two equations imply that $\overline{K}^{\diamond}(T) \leq \overline{K}(T)$, which contradicts the definition of T.

Case 2. $\overline{X}(T) \geq 1$. In this case, (3.7) shows that $\langle \mathbf{1}, \overline{\nu}_T \rangle = 1$. Since the pairs $(\overline{\nu}, \overline{K})$ and $(\overline{\nu}^{\diamond}, \overline{K}^{\diamond})$ both satisfy the fluid equation (3.5), Corollary 4.4 and (3.8) show that

$$\overline{D}(T) = \int_{[0,M)} \frac{G(x+T) - G(x)}{1 - G(x)} \overline{\nu}_0(dx) + \int_0^T g(T-s)\overline{K}(s)\, ds. \tag{4.14}$$

If $G(T) > 0$, then by the definition of T,

$$\int_0^T g(T-s)\overline{K}(s)\, ds > \int_0^T g(T-s)\big(\overline{K}^{\diamond}(s) - \varepsilon\big)\, ds.$$

Together with (4.14), the corresponding equation for $\overline{D}^{\diamond}$ and the fact that $\overline{\nu}_0 = \overline{\nu}_0^{\diamond}$, this shows that

$$\overline{D}(T) > \overline{D}^{\diamond}(T) - \varepsilon G(T) \geq \overline{D}^{\diamond}(T) - \varepsilon.$$

On the other hand, if $G(T) = 0$, then (4.14) implies that

$$\overline{D}(T) = \overline{D}^{\diamond}(T) > \overline{D}^{\diamond}(T) - \varepsilon.$$

Combining the last two inequalities with (3.8) and the case assumption, we obtain

$$\begin{aligned} \overline{K}(T) - \overline{K}^{\diamond}(T) &= \langle \mathbf{1}, \overline{\nu}_T \rangle - \langle \mathbf{1}, \overline{\nu}_T^{\diamond} \rangle + \overline{D}(T) - \overline{D}^{\diamond}(T) \\ &= 1 - \langle \mathbf{1}, \overline{\nu}_T^{\diamond} \rangle + \overline{D}(T) - \overline{D}^{\diamond}(T) \\ &> -\varepsilon, \end{aligned}$$

which again contradicts the definition of T.

Thus we have shown that $T = \infty$ or, equivalently, that $\overline{K}(t) \geq \overline{K}^{\diamond}(t) - \varepsilon$ for every $t \in [0, \infty)$ and $\varepsilon > 0$. Sending $\varepsilon \to 0$, we conclude that $\overline{K}(t) \geq \overline{K}^{\diamond}(t)$ for

$t \in [0, \infty)$. Together with (3.6), Corollary 4.4 and the fact that $\overline{\nu}_0^\diamond = \overline{\nu}_0$, this implies that for every $t \in [0, \infty)$, we have

$$\overline{D}(t) - \overline{D}^\diamond(t) = \int_0^t g(t-s)\big(\overline{K}(s) - \overline{K}^\diamond(s)\big)\, ds \geq 0,$$

which concludes the proof of the lemma. □

REMARK 4.8. A similar maximality property (in terms of a stochastic, rather than pathwise, ordering) is satisfied by the "pre-limit" processes describing $G/GI/N$ queues (see, e.g., [1], Theorem 1.2 of Chapter XII). It is also worthwhile to note the connection between Lemma 4.7 and a minimality property associated with the one-dimensional reflection map that is used to characterize single-server queues. In the latter case, the so-called complementarity condition plays the role of the nonidling condition here, and ensures minimality of the associated constraining term (see, e.g., [9]).

4.3. *Analysis of the age equation* (4.2). The goal of this section is to establish Theorem 4.1. In fact, we establish the somewhat more general result of identifying solutions to the so-called abstract age equation (see Definition 4.9 and Corollary 4.17 below). The abstract age equation and a related integral equation, which we refer to as the simplified age equation, are first introduced in Section 4.3.1. The simplified age equation is shown to have a unique and explicit solution in Section 4.3.2. Using a simple correspondence between solutions of the abstract age equation and those of the simplified age equation, an explicit representation for the unique solution to the abstract age equation is then obtained in Section 4.3.3. These results are then combined in Section 4.3.4 to establish Theorem 4.1.

Throughout the analysis, the characterization of Radon measures described in Section 1.2.2 is repeatedly used, often without explicit mention. For conciseness, the following notations are also used. Let $\tilde{\mathcal{M}}$ be the space of finite Radon measures on $\mathbb{R}^2$ whose support lies in $[0, M) \times \mathbb{R}_+$, and let $\tilde{\mathcal{C}}$ be the space of continuous functions on $\mathbb{R}^2$ with compact support in $[0, M) \times \mathbb{R}_+$. Also, let $\tilde{\mathcal{C}}^{1,1}$ be the subset of functions φ in $\tilde{\mathcal{C}}$ for which the directional derivative $\varphi_x + \varphi_s$ exists and is continuous. The integral with respect to any Radon measure ζ on $\mathbb{R}^2$ is denoted by

$$\zeta(\varphi) \doteq \int\!\!\int_{\mathbb{R}^2} \varphi(x, s)\zeta(dx, ds), \qquad \varphi \in \mathcal{C}_c(\mathbb{R}^2).$$

As in the rest of the paper, given a measure θ on $[0, M)$, and a θ-integrable function f on $[0, \infty)$, the integral of f with respect to θ over $[0, M)$ is denoted by $\langle f, \theta \rangle$. Lebesgue measure on $\mathbb{R}^2$ is denoted by σ, and, for $m_1, m_2 \in [0, \infty)$, the corresponding rectangle is represented by

$$\mathcal{R}_{m_1,m_2} \doteq [-m_1, m_1] \times [-m_2, m_2]. \tag{4.15}$$

Given a Radon measure ζ on $\mathbb{R}^2$ and a function $f \in \mathcal{C}_c^\infty(\mathbb{R}^2)$, recall that the convolution $f \star \zeta$ is the absolutely continuous measure whose density (with respect to Lebesgue measure) lies in $C_c^\infty(\mathbb{R}^2)$ and is given explicitly by

$$\frac{d(f \star \zeta)}{d\sigma}(y,u) = \int\!\!\int_{\mathbb{R}^2} f(y-x, u-s)\zeta(dx, ds). \tag{4.16}$$

Definitions and standard properties of convolutions can be found in Section 2.5.9 of [21].

4.3.1. *The abstract and simplified age equations.* We first introduce the abstract age equation.

DEFINITION 4.9 (Abstract age equation). Given $\gamma \in \tilde{\mathcal{M}}$ and $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$, $\{\zeta_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ is said to solve the abstract age equation for γ and ℓ if and only if the measure $\ell\zeta$, defined by

$$(\ell\zeta)(\varphi) \doteq \int_0^\infty \langle \ell(\cdot)\varphi(\cdot, s), \zeta_s\rangle\, ds, \qquad \varphi \in \tilde{\mathcal{C}},$$

is a well defined measure that belongs to $\tilde{\mathcal{M}}$, and for every $\varphi \in \tilde{\mathcal{C}}^{1,1}$,

$$-\int_0^\infty \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \zeta_s\rangle\, ds = -(\ell\zeta)(\varphi) + \gamma(\varphi). \tag{4.17}$$

In order to analyze the abstract age equation, we will find it convenient to first study a related, but somewhat simpler, integral equation, which we refer to as the simplified age equation.

DEFINITION 4.10 (Simplified age equation). Given $\tilde{\gamma} \in \tilde{\mathcal{M}}$, $\{\mu_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ is said to solve the simplified age equation for $\tilde{\gamma}$ if and only if for every $\tilde{\varphi} \in \tilde{\mathcal{C}}^{1,1}$,

$$-\int_0^\infty \langle \tilde{\varphi}_x(\cdot, s) + \tilde{\varphi}_s(\cdot, s), \mu_s\rangle\, ds = \tilde{\gamma}(\tilde{\varphi}). \tag{4.18}$$

REMARK 4.11. It follows immediately from the definitions that any $\{\zeta_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ satisfies the abstract age equation for $\gamma \in \tilde{\mathcal{M}}$ and $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$ if and only if $\{\zeta_t\}_{t\geq 0}$ satisfies the simplified age equation for $\tilde{\gamma} = \gamma - \ell\zeta$.

Recall the hazard rate function $h \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$, the Radon measure υ_0 on $[0, M)$ and the function Z that has finite variation on every bounded interval, which were introduced in Section 4.1. Given $(m_1, m_2) \in [0, M) \times [0,\infty)$, let $|\upsilon_0|_{\mathrm{TV},m_1}$ denote the total variation of the Radon measure υ_0 on $[0, m_1]$, and let $\mathcal{V}\mathrm{ar}(Z; [0, m_2])$ denote the total variation of the function Z on the interval $[0, m_2]$. We now introduce some definitions that will help elucidate the connection between the abstract and

simplified age equations introduced above, and the age equation (4.2) associated with h, υ_0 and Z. Consider the measure $\xi = \xi(\upsilon_0, Z)$ on $\mathbb{R}^2$ defined by

$$\xi(\varphi) \doteq \int_{[0,M)} \varphi(x,0)\upsilon_0(dx) + \int_{[0,\infty)} \varphi(0,s)\,dZ(s), \qquad \varphi \in \mathcal{C}_c(\mathbb{R}^2). \tag{4.19}$$

Clearly, for all $\varphi \in \mathcal{C}_c(\mathbb{R}^2)$ such that $\operatorname{supp}(\varphi) \subseteq \mathcal{R}_{m_1,m_2}$, $\xi(\varphi)$ satisfies

$$|\xi(\varphi)| \le \|\varphi\|_\infty\big(|\upsilon_0|_{\mathrm{TV},m_1} + \mathcal{V}\mathrm{ar}(Z;[0,m_2])\big). \tag{4.20}$$

Moreover, $\xi(\varphi) = 0$ for all φ such that $\operatorname{supp}(\varphi) \cap [0,M) \times \mathbb{R}_+ = \varnothing$. Therefore, ξ is a Radon measure on $\mathbb{R}^2$ that has support in $[0,M) \times [0,\infty)$ and, hence, lies in $\tilde{\mathcal{M}}$.

Now, suppose that $\{\overline{\nu}_t\}_{t\ge 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ satisfies condition (4.1), and let $h\overline{\nu}$ be the measure on $\mathbb{R}^2$ defined by

$$(h\overline{\nu})(\varphi) \doteq \int_0^\infty \langle h(\cdot)\varphi(\cdot,s), \overline{\nu}_s\rangle\,ds, \qquad \varphi \in \mathcal{C}_c(\mathbb{R}^2). \tag{4.21}$$

Then (4.1) shows that $h\overline{\nu}$ is a Radon measure, and it is clear from (4.21) that $h\overline{\nu}$ has support in $[0,M) \times \mathbb{R}_+$. Therefore, $h\overline{\nu} \in \tilde{\mathcal{M}}$. Also, define $\xi^{\overline{\nu}} = \xi^{\overline{\nu}}(\upsilon_0, Z)$ by

$$\xi^{\overline{\nu}} \doteq \xi - h\overline{\nu}, \tag{4.22}$$

with ξ as defined in (4.19). Clearly, $\xi^{\overline{\nu}}$ also lies in $\tilde{\mathcal{M}}$. We now derive some alternative characterizations of solutions to the age equation associated with υ_0, Z and h.

LEMMA 4.12. *Suppose* $\{\overline{\nu}_t\}_{t\ge 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ *satisfies* (4.1). *Then, for* $\upsilon_0 \in \mathcal{M}[0,M)$ *and* $Z \in \mathcal{B}V_0[0,\infty)$, *the following statements are equivalent*:

(1) $\{\overline{\nu}_t\}_{t\ge 0}$ *satisfies the age equation* (4.2) *for* υ_0 *and* Z;
(2) $\{\overline{\nu}_t\}_{t\ge 0}$ *satisfies the abstract age equation* (4.17) *for* $\xi = \xi(\upsilon_0, Z)$ *and* h;
(3) $\{\overline{\nu}_t\}_{t\ge 0}$ *satisfies the simplified age equation* (4.18) *for* $\xi^{\overline{\nu}} = \xi^{\overline{\nu}}(\upsilon_0, Z)$.

PROOF. Fix $\{\overline{\nu}_t\}_{t\ge 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ that satisfies (4.1), and let $h\overline{\nu} \in \tilde{\mathcal{M}}$ be the Radon measure defined above in (4.21). We first show that (1) implies (2). Suppose $\{\overline{\nu}_t\}_{t\ge 0}$ satisfies the age equation (4.2) for all $\varphi \in \tilde{\mathcal{C}}^{1,1}$ and $t \in [0,\infty)$. Then, because φ has compact support in $[0,M) \times \mathbb{R}_+$, for all sufficiently large t, the left-hand side of (4.2) equals zero. Therefore, on sending $t \to \infty$ in (4.2), a little rearrangement shows that

$$\begin{aligned} &-\int_0^\infty \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \overline{\nu}_s\rangle\,ds \\ &\quad = -\int_0^\infty \langle h(\cdot)\varphi(\cdot,s), \overline{\nu}_s\rangle\,ds + \langle \varphi(\cdot,0), \upsilon_0\rangle \\ &\qquad + \int_{[0,\infty)} \varphi(0,s)\,dZ(s) \\ &\quad = -h\overline{\nu}(\varphi) + \xi(\varphi), \end{aligned} \tag{4.23}$$

which shows that $\{\overline{\nu}_t\}_{t\geq 0}$ satisfies the abstract age equation for ξ and h.

The equivalence of properties (2) and (3) is an immediate consequence of the definitions (also see Remark 4.11). Therefore, to complete the proof, it suffices to show that (3) implies (1). Suppose that $\{\overline{\nu}_t\}_{t\geq 0}$ satisfies the simplified age equation for $\xi^{\overline{\nu}}$. Then, in particular, (4.23) holds for every $\varphi \in \tilde{\mathcal{C}}^{1,1}$. A standard mollification argument will now be used to show that then $\{\overline{\nu}_t\}_{t\geq 0}$ satisfies the age equation for υ_0 and Z. Fix $t \in [0,\infty)$ that is a continuity point of $\{\overline{\nu}_t\}_{t\geq 0}$, and let $\{c_n\} = \{c_n^t\}$ be a uniformly bounded sequence of functions in $\mathcal{C}^\infty(\mathbb{R})$ such that the negative of their derivatives, $-c_n'$, are probability density functions on $\mathbb{R}$ and, as $n \to \infty$, $c_n(s) \to \mathbb{1}_{[0,t]}(s)$ and the sequence of probability measures $-c_n'(s)\,ds$ converge weakly to the Dirac measure concentrated at t. (For instance, consider $c_n(s) = \int_s^\infty n\rho(n(u-t))\,du$, $s \in [0,\infty)$, where $\rho(x) = k\exp(1/((x-1)^2-1))\mathbb{1}_{[0,2]}(x)$ and k is the appropriate normalization constant that makes ρ a probability density.) Given $\varphi \in \tilde{\mathcal{C}}^{1,1}$, define $\tilde{\varphi}_n(x,s) \doteq \varphi(x,s)c_n(s)$ for all $(x,s) \in \mathbb{R}^2$, and note that $\tilde{\varphi}_n \in \tilde{\mathcal{C}}^{1,1}$. Replacing φ by $\tilde{\varphi}_n$ in (4.23) then yields, for every $n \in \mathbb{N}$,

$$\begin{aligned}\int_0^\infty -c_n'(s)\langle\varphi(\cdot,s),\overline{\nu}_s\rangle\,ds - \int_0^\infty c_n(s)\langle\varphi_x(\cdot,s)+\varphi_s(\cdot,s),\overline{\nu}_s\rangle\,ds \\ = -\int_0^\infty c_n(s)\langle h(\cdot)\varphi(\cdot,s),\overline{\nu}_s\rangle\,ds + c_n(0)\langle\varphi(\cdot,0),\upsilon_0\rangle \\ + \int_{[0,\infty)} c_n(s)\varphi(0,s)\,dZ(s).\end{aligned}$$

Now, take limits as $n \to \infty$ in the above equation. Since $\operatorname{supp}(\varphi) \subset [0,M) \times [0,T]$ for some $T < \infty$, the right-continuity of $\{\overline{\nu}_t\}_{t\geq 0}$ implies $s \mapsto \langle\varphi(\cdot,s),\overline{\nu}_s\rangle$ is uniformly bounded. The weak convergence $-c_n'(s)\,ds \xrightarrow{w} \delta_t$ and the fact that t is a continuity point for $s \mapsto \langle\varphi(\cdot,s),\overline{\nu}_s\rangle$ then shows that the first term above converges to $\langle\varphi(\cdot,t),\overline{\nu}_t\rangle$. The limit of the remaining terms can be obtained using the fact that $c_n \to \mathbb{1}_{[0,t]}$ and the dominated convergence theorem [whose application is justified by the inequality (4.1) and the uniform boundedness of the sequence of functions $c_n, n \in \mathbb{N}$] to yield

$$\begin{aligned}\langle\varphi(\cdot,t),\overline{\nu}_t\rangle - \int_0^t \langle\varphi_x(\cdot,s)+\varphi_s(\cdot,s),\overline{\nu}_s\rangle\,ds \\ = -\int_0^t \langle h(\cdot)\varphi(\cdot,s),\overline{\nu}_s\rangle\,ds + \langle\varphi(\cdot,0),\upsilon_0\rangle + \int_{[0,t]}\varphi(0,s)\,dZ(s).\end{aligned}$$

This shows that $\{\overline{\nu}_t\}_{t\geq 0}$ satisfies the age equation (4.2) associated with υ_0 and Z, and so property (1) follows. □

Since $\xi^{\overline{\nu}}$ depends on $\{\overline{\nu}_t\}_{t\geq 0}$, Lemma 4.12 only shows that solutions $\{\overline{\nu}_t\}_{t\geq 0}$ to the age equation satisfy the simplified age equation in an implicit sense. Nevertheless, this property is used in the proof of Theorem 4.1 in Section 4.3.4 in order to justify the application of the estimate obtained in Proposition 4.15 below, and the

application of the more explicit correspondence result obtained in Proposition 4.16 to $\{\overline{\nu}_t\}_{t\geq 0}$.

4.3.2. *Solution to the simplified age equation.* In Lemma 4.13 below, it is shown that the solution to the simplified age equation is unique and can be represented in terms of the following maps. For $t \geq 0$, consider the map Λ^t that takes $f \in \mathcal{C}_c(\mathbb{R})$ to the measurable function Λ^t_f defined by

$$\Lambda^t_f(x,s) \doteq f(x+t-s)\mathbb{1}_{[0,t]}(s), \qquad (x,s) \in \mathbb{R}^2. \tag{4.24}$$

Observe that for any $t > 0$ and $f \in \mathcal{C}_c(\mathbb{R})$,

$$\|\Lambda^t_f\|_\infty \leq \|f\|_\infty \tag{4.25}$$

and

$$\text{supp}(f) \subseteq [-\tilde{m}, m] \quad \Rightarrow \quad \text{supp}(\Lambda^t_f) \subseteq [-\tilde{m}-t, m] \times [0,t]. \tag{4.26}$$

Also, let the map $\pi : \mathcal{C}_c(\mathbb{R}^2) \mapsto \mathcal{C}(\mathbb{R}^2)$ that maps φ to π_φ, be defined by

$$\pi_\varphi(x,s) \doteq \int_0^\infty \varphi(x+r, s+r)\, dr, \qquad (x,s) \in \mathbb{R}^2. \tag{4.27}$$

It is easily verified that for any $\varphi \in \mathcal{C}_c(\mathbb{R}^2)$ with $\text{supp}(\varphi) \subseteq \mathcal{R}_{m_1,m_2}$,

$$\begin{aligned} &\|\pi_\varphi\|_\infty \leq 2\sqrt{m_1^2+m_2^2}\|\varphi\|_\infty \quad \text{and} \\ &\text{supp}(\pi_\varphi) \subseteq (-\infty, m_1] \times (-\infty, m_2]. \end{aligned} \tag{4.28}$$

LEMMA 4.13. *The simplified age equation associated with $\tilde{\gamma} \in \tilde{\mathcal{M}}$ has a unique solution $\{\mu_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ that is given explicitly by*

$$\langle f, \mu_t\rangle = \tilde{\gamma}(\Lambda^t_f), \qquad f \in \mathcal{C}_c[0,M), t \geq 0. \tag{4.29}$$

Moreover, for every $\varphi \in \tilde{\mathcal{C}}$,

$$\int_0^\infty \langle \varphi(\cdot,t), \mu_t\rangle\, dt = \tilde{\gamma}(\pi_\varphi). \tag{4.30}$$

PROOF. Let $\{\mu_t\}_{t\geq 0}$ be as defined in (4.29). Then, to show that $\{\mu_t\}_{t\geq 0}$ belongs to $\mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$, it clearly suffices to show that for every $t \geq 0$ and $f \in \mathcal{C}_c[0,M)$, $\tilde{\gamma}(\Lambda^{t+\varepsilon}_f) \to \tilde{\gamma}(\Lambda^t_f)$ as $\varepsilon \to 0$. However, the latter limit holds due to the pointwise convergence $\Lambda^{t+\varepsilon}_f \to \Lambda^t_f$ and the dominated convergence theorem, whose application is justified by the properties in (4.25) and (4.26). Next, to show that $\{\mu_t\}_{t\geq 0}$ satisfies the simplified age equation (4.18), we first claim that (4.29) implies (4.30). Given $\varphi \in \tilde{\mathcal{C}}$, note that

$$\begin{aligned} \left(\int_0^\infty \Lambda^t_{\varphi(\cdot,t)}\, dt\right)(x,s) &= \int_0^\infty \varphi(x+t-s,t)\mathbb{1}_{[s,\infty)}(t)\, dt \\ &= \int_0^\infty \varphi(x+r, s+r)\, dr = \pi_\varphi(x,s). \end{aligned}$$

Therefore, first replacing f in (4.29) by $\varphi(\cdot, t)$, then integrating both sides of (4.29) over $t \in [0, \infty)$ and using Fubini's theorem, we see that

$$\int_0^\infty \langle \varphi(\cdot, t), \mu_t \rangle \, dt = \int_0^\infty \tilde{\gamma}\big(\Lambda^t_{\varphi(\cdot,t)}\big) \, dt = \tilde{\gamma}\bigg(\int_0^\infty \Lambda^t_{\varphi(\cdot,t)} \, dt\bigg) = \tilde{\gamma}(\pi_\varphi),$$

which proves the claim. Then, for $\tilde{\varphi} \in \tilde{\mathcal{C}}^{1,1}$, replacing φ in (4.30) by $\tilde{\varphi}_x + \tilde{\varphi}_s$ and observing that $\pi_{\tilde{\varphi}_x + \tilde{\varphi}_s} = -\tilde{\varphi}$ (this uses the fact that $\tilde{\varphi}$ has compact support), it follows that $\{\mu_t\}_{t \geq 0}$ satisfies the simplified age equation.

It only remains to establish uniqueness. Let $\{\mu^i_t\}_{t \geq 0}$, $i = 1, 2$, be two solutions to the simplified age equation for $\tilde{\gamma}$, and define $\eta_t \doteq \mu^1_t - \mu^2_t$, $t \geq 0$. Then η_0 is the zero measure and for every $\tilde{\varphi} \in \tilde{\mathcal{C}}^{1,1}$,

$$\int_0^\infty \langle \tilde{\varphi}_x(\cdot, s) + \tilde{\varphi}_s(\cdot, s), \eta_s \rangle \, ds = 0. \tag{4.31}$$

Fix $\varphi \in \tilde{\mathcal{C}}^{1,1}$, and define $\tilde{\varphi} \doteq \pi_\varphi$. Then $\operatorname{supp}(\tilde{\varphi}) \cap ([0, M) \times [0, \infty))$ is compact by (4.28), $\tilde{\varphi}$ lies in $\tilde{\mathcal{C}}^{1,1}$ and $\tilde{\varphi}_x + \tilde{\varphi}_s = (\pi_\varphi)_x + (\pi_\varphi)_s = -\varphi$. When substituted into (4.31) this shows that

$$\int_0^\infty \langle \varphi(\cdot, s), \eta_s \rangle \, ds = 0, \qquad \varphi \in \tilde{\mathcal{C}}^{1,1}.$$

Standard approximation arguments can now be used to show that η is identically zero. Specifically, let $\{\rho^n\}_{n \in \mathbb{N}}$ be a sequence of mollifiers, that is, nonnegative functions in $\mathcal{C}^\infty_c(\mathbb{R})$, with ρ^n having support in $[0, 1/n]$ and $\int_{\mathbb{R}} \rho^n(x) \, dx = 1$ and such that, as $n \to \infty$, the family of measures $\rho^n(x) \, dx$ converge vaguely to the delta distribution δ_0. For any $t > 0$ that is a continuity point of $\{\eta_s\}_{s \geq 0}$ and $f \in \mathcal{C}^1_c[0, M)$, first replace φ in the last display by $\varphi_n(x, s) \doteq f(x)\rho^n(t - s)$ then take limits as $n \to \infty$ and use the right continuity of the function $s \mapsto \langle f, \eta_s \rangle$ at t and the vague convergence of ρ^n to δ_0 in order to conclude that $\langle f, \eta_t \rangle = 0$. Since $\mathcal{C}^1_c[0, M)$ is a determining class for Radon measures on $[0, M)$, it follows that each η_t is identically zero for every t that is a continuity point of $\{\eta_t\}_{t \geq 0}$. The right continuity of $\{\eta_t\}_{t \geq 0}$ then implies that η_t is identically zero for every $t \geq 0$, and so uniqueness follows. □

Replacing $\tilde{\gamma}$ in Lemma 4.13 by the measure ξ defined in (4.19), we obtain the following result.

COROLLARY 4.14. *Given $\upsilon_0 \in \mathcal{M}[0, M)$ and $Z \in \mathcal{BV}_0[0, \infty)$, the unique solution $\{\mu_t\}_{t \geq 0}$ to the simplified age equation associated with $\xi = \xi(\upsilon_0, Z)$ satisfies*

$$\langle f, \mu_t \rangle = \langle f(\cdot + t), \upsilon_0 \rangle + \int_{[0,t]} f(t - s) \, dZ(s). \tag{4.32}$$

We now establish a property of solutions to simplified age equations that will be used in the proof of the correspondence property in Section 4.3.3. Given any $\{\mu_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$, consider the measure $\ell\mu$ defined by

$$(\ell\mu)(\varphi) \doteq \int_0^\infty \left(\int_{[0,M)} \ell(x)\varphi(x,s)\mu_t(dx) \right) ds, \qquad \varphi \in \tilde{\mathcal{C}}. \tag{4.33}$$

When ℓ is continuous, $\varphi \in \tilde{\mathcal{C}}$ implies $\ell\varphi \in \tilde{\mathcal{C}}$, and hence $\ell\mu$ is a well-defined Radon measure that lies in $\tilde{\mathcal{M}}$. However, when $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0,M)$, $\ell\mu$ need not always be well defined for arbitrary $\{\mu_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$. In Proposition 4.15 below, we show that if $\{\mu_t\}_{t\geq 0}$ satisfies the simplified age equation, then $\ell\mu$ is a well-defined Radon measure for any ℓ that is locally integrable on $[0,M)$. A real-valued function $\tilde{L}$ on $[0,M)\times[0,\infty)$ is said to be coordinate-wise increasing if for every $(x,t) \in [0,M)$, $\tilde{L}(\cdot,t)$ and $\tilde{L}(x,\cdot)$ are increasing functions on $[0,M)$ and $[0,\infty)$, respectively.

PROPOSITION 4.15. *Suppose $\{\mu_t\}_{t\geq 0}$ solves the simplified age equation for some $\tilde{\gamma} \in \tilde{\mathcal{M}}$. Then there exists a coordinate-wise increasing function $\tilde{L}$ on $[0,M)\times[0,\infty)$ such that given any $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0,\infty)$, for every $m_1 \in [0,M)$, $m_2 \in (0,\infty)$ and $\varphi \in \mathcal{C}_c([0,M)\times\mathbb{R}_+)$ with* $\mathrm{supp}(\varphi) \subseteq \mathcal{R}_{m_1,m_2}$,

$$\begin{aligned} &\left| \int_0^\infty \left(\int_{[0,M)} \ell(x)\varphi(x,s)\mu_t(dx) \right) ds \right| \\ &\qquad \leq \left(\int_0^{m_1} |\ell(x)|\, dx \right) \tilde{L}(m_1,m_2)\|\varphi\|_\infty. \end{aligned} \tag{4.34}$$

Consequently, the definition in (4.33) yields a well-defined Radon measure $\ell\mu$ that belongs to $\tilde{\mathcal{M}}$.

PROOF. We first establish (4.34) for continuous ℓ. Fix $\tilde{\gamma}$ and $\{\mu_t\}_{t\geq 0}$ as in the statement of the proposition, and let $L : [0,M)\times\mathbb{R}_+ \mapsto \mathbb{R}_+$ be the component-wise nondecreasing function such that

$$|\tilde{\gamma}(\varphi)| \leq L(m_1,m_2)\|\varphi\|_\infty \qquad \forall \varphi \in \mathcal{C}_c(\mathbb{R}^2) \text{ with } \mathrm{supp}(\varphi) \subset \mathcal{R}_{m_1,m_2}. \tag{4.35}$$

Such a function L exists since $\tilde{\gamma}$ is, by assumption, a Radon measure (see Section 1.2.2). Let μ be the measure on $\mathbb{R}^2$ defined by

$$\mu(\varphi) \doteq \int_0^\infty \left(\int_{[0,M)} \varphi(x,s)\mu_s(dx) \right) ds, \qquad \varphi \in \tilde{\mathcal{C}}. \tag{4.36}$$

Since $\{\mu_t\}_{t\geq 0}$ lies in $\mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$, (4.36) shows that μ belongs to $\tilde{\mathcal{M}}$. (Note that we will always write μ for the Radon measure on $\mathbb{R}^2$ and $\{\mu_t\}_{t\geq 0}$ for the measure-valued function in order to keep these two quantities distinct.) The basic idea behind the proof (for continuous ℓ) is to construct a sequence $\mu^n, n \in \mathbb{N}$,

of Radon measures on $\mathbb{R}^2$ such that for each $n \in \mathbb{N}$, μ^n is absolutely continuous with respect to Lebesgue measure and satisfies the following two properties: (i) for every $m_1 \in [0, M)$, $m_2 \in (0, \infty)$, for all n sufficiently large such that $m_1 + 1/n < M$, and all $y < m_1$,

$$\left| \int_0^{m_2} \frac{d\mu^n}{d\sigma}(y, u)\, du \right| \le L\left(m_1 + \frac{1}{n}, m_2 + \frac{1}{n}\right), \tag{4.37}$$

and (ii) for continuous ℓ, as $n \to \infty$,

$$\mu^n(\ell\varphi) \to \mu(\ell\varphi) = \ell\mu(\varphi), \qquad \varphi \in \tilde{\mathcal{C}}. \tag{4.38}$$

Given such a sequence, for any $\varphi \in \tilde{\mathcal{C}}$, with $\operatorname{supp}(\varphi) \subseteq \mathcal{R}_{m_1,m_2}$,

$$\mu^n(\ell\varphi) = \int_0^{m_1} \left(\int_0^{m_2} \frac{d\mu^n}{d\sigma}(x, u)\varphi(x, u)\, du \right) \ell(x)\, dx, \qquad n \in \mathbb{N}.$$

Together with the estimate in (4.37), this shows that for all $n \in \mathbb{N}$ sufficiently large such that $m_1 + 1/n < M$,

$$|\mu^n(\ell\varphi)| \le \left(\int_0^{m_1} \ell(x)\, dx \right) \|\varphi\|_\infty L\left(m_1 + \frac{1}{n}, m_2 + \frac{1}{n}\right).$$

Taking limits as $n \to \infty$ and using (4.38), we obtain (4.34) with $\tilde{L}(m_1, m_2) \doteq L(m_1+, m_2+)$ for $\varphi \in \tilde{\mathcal{C}}$ such that $\operatorname{supp}(\varphi) \subset \mathcal{R}_{m_1,m_2}$. This implies that (4.34) holds for all $\varphi \in \tilde{\mathcal{C}}$ when ℓ is continuous.

We now construct an approximating sequence $\mu^n, n \in \mathbb{N}$, that satisfies properties (4.37) and (4.38) mentioned above. Let $\{\rho^n\}_{n\in\mathbb{N}}$ be a sequence of mollifiers, where for each $n \in \mathbb{N}$, ρ^n is a nonnegative function in $\mathcal{C}_c^\infty(\mathbb{R}^2)$ with support in $\mathcal{R}_{1/n,1/n}$ that has integral 1 and, as $n \to \infty$, converges vaguely to the delta distribution $\delta_{(0,0)}$, defined by $\delta_{(0,0)}(\varphi) = \varphi(0, 0)$. For each $n \in \mathbb{N}$, define μ^n to be the convolution $\rho^n \star \mu$. In other words (see the discussion of convolutions at the beginning of Section 4.3), μ^n is absolutely continuous with respect to Lebesgue measure σ on $\mathbb{R}^2$, with density $d\mu^n/d\sigma$ in $\mathcal{C}_c^\infty(\mathbb{R}^2)$ that has the explicit form

$$\begin{aligned} \frac{d\mu^n}{d\sigma}(y, u) &= \int\!\!\int_{\mathbb{R}^2} \rho^n(y - x, u - s)\mu(dx, ds) \\ &= \int_0^\infty \left(\int_{[0,M)} \rho^n(y - x, u - s)\mu_s(dx) \right) ds. \end{aligned} \tag{4.39}$$

Then, for any $m_2 \in (0, \infty)$, by Fubini's theorem,

$$\int_0^{m_2} \frac{d\mu^n}{d\sigma}(y, u)\, du = \int_0^\infty \left(\int_{[0,M)} \theta_n^y(x, s)\mu_s(dx) \right) ds, \tag{4.40}$$

where, for $y \in \mathbb{R}$,

$$\theta_n^y(x, s) \doteq \int_0^{m_2} \rho^n(y - x, u - s)\, du, \qquad (x, s) \in [0, M) \times [0, \infty).$$

Clearly, θ_n^y is continuous and, for any $m_1 \in [0, M)$ and $y \in [0, m_1]$, $\mathrm{supp}(\theta_n^y) \subseteq \mathcal{R}_{m_1+1/n, m_2+1/n}$. Therefore, for all sufficiently large n (such that $m_1 + 1/n < M$), we have $\theta_n^y \in \tilde{\mathcal{C}}$. Since $\{\mu_t\}_{t\geq 0}$ solves the simplifed age equation for $\tilde{\gamma}$, relation (4.30) of Lemma 4.13 can be invoked to rewrite the right-hand side of (4.40) so as to obtain

$$\int_0^{m_2} \frac{d\mu^n}{d\sigma}(y,u)\,du = \tilde{\gamma}(\pi_{\theta_n^y}). \tag{4.41}$$

From the definition of π in (4.27) and the expression for θ_n^y given above, we have

$$\pi_{\theta_n^y}(x,s) = \int_0^\infty \left(\int_0^{m_2} \rho^n(y-x-r, u-s-r)\,du \right) dr.$$

Since $\mathrm{supp}(\rho^n) \subseteq \mathcal{R}_{1/n,1/n}$ and ρ^n is nonnegative with integral over $\mathbb{R}^2$ equal to 1, it follows that $\|\pi_{\theta_n^y}\|_\infty \leq 1$ and

$$\mathrm{supp}(\pi_{\theta_n^y}) \cap \mathbb{R}_+^2 \subset \mathcal{R}_{m_1+1/n, m_2+1/n}.$$

Due to (4.35) and the fact that $\mathrm{supp}(\tilde{\gamma}) \subset \mathbb{R}_+^2$, this then implies that

$$|\tilde{\gamma}(\pi_{\theta_n^y})| \leq L\left(m_1 + \frac{1}{n}, m_2 + \frac{1}{n}\right). \tag{4.42}$$

When combined with (4.41), this yields (4.37).

Next, we show that (4.38) holds for continuous ℓ. For every $\tilde{\varphi} \in \tilde{\mathcal{C}}$, multiplying both sides of (4.39) by $\tilde{\varphi}(y,u)$, then integrating over $(y,u) \in \mathbb{R}^2$ and using Fubini's theorem, we see that

$$\mu^n(\tilde{\varphi}) = \int_0^\infty \left(\int_{[0,M)} (\check{\rho}_n \star \tilde{\varphi})(x,s)\mu_s(dx) \right) ds = \mu(\check{\rho}_n \star \tilde{\varphi}), \tag{4.43}$$

where $\check{\rho}^n(x,s) \doteq \rho^n(-x,-s)$ for $(x,s) \in \mathbb{R}^2$. Now, fix $m_1 \in [0, M)$ and $m_2 < \infty$. If $\mathrm{supp}(\tilde{\varphi}) \subseteq \mathcal{R}_{m_1,m_2}$ then, since $\mathrm{supp}(\check{\rho}^n) \subseteq \mathcal{R}_{1/n,1/n}$, it follows that $\mathrm{supp}(\check{\rho}^n \star \tilde{\varphi}) \subseteq \mathcal{R}_{m_1+1/n, m_2+1/n}$. Therefore, for all n sufficiently large so that $m_1 + 1/n < M$, $\check{\rho}^n \star \tilde{\varphi} \in \tilde{\mathcal{C}}$. Since $\{\mu_t\}_{t\geq 0}$ satisfies the simplified age equation, (4.43) and the relation (4.30) of Lemma 4.13 show that for all n such that $m_1 + 1/n < M$,

$$\mu^n(\tilde{\varphi}) = \tilde{\gamma}(\pi_{\check{\rho}^n \star \tilde{\varphi}}), \qquad \tilde{\varphi} \in \tilde{\mathcal{C}} \qquad \text{with } \mathrm{supp}(\tilde{\varphi}) \subseteq \mathcal{R}_{m_1,m_2}. \tag{4.44}$$

Next, send $n \to \infty$ in (4.44). Using the fact that for all sufficiently large n, the functions $\pi_{\check{\rho}^n \star \varphi}$ are uniformly bounded and have common compact support in $[0, M) \times \mathbb{R}_+$, $\check{\rho}^n \star \tilde{\varphi} \to \tilde{\varphi}$ and, hence, $\pi_{\check{\rho}^n \star \tilde{\varphi}} \to \pi_{\tilde{\varphi}}$ pointwise, we apply the dominated convergence theorem to conclude that

$$\mu^n(\tilde{\varphi}) = \tilde{\gamma}(\pi_{\check{\rho}^n \star \tilde{\varphi}}) \to \tilde{\gamma}(\pi_{\tilde{\varphi}}) = \mu(\tilde{\varphi}),$$

where the last equality holds due to Lemma 4.13 because $\{\mu_t\}_{t\geq 0}$ satisfies the simplified age equation for $\tilde{\gamma}$. In turn, $\varphi \in \tilde{\mathcal{C}}$ implies $\ell\varphi \in \tilde{\mathcal{C}}$ because ℓ is continuous. Therefore, we can replace φ by $\ell\varphi$ in the last display to conclude that (4.38) holds.

Thus, we have constructed a sequence $\{\mu^n\}_{n\in\mathbb{N}}$ of Radon measures on $\mathbb{R}^2$ that satisfies (4.37) and (4.38). Therefore, by the argument given in the first paragraph of the proof, it follows that (4.34) holds for continuous ℓ. Let $\ell\mu$ be the Radon measure for which $\ell\mu(\varphi)$ equals the left-hand side of (4.34). Then, the estimate (4.34) implies that the product mapping $\ell \mapsto \ell\mu$ from $\mathcal{C}[0, M) \subset \mathcal{L}^1_{\mathrm{loc}}[0, M)$ to $\mathcal{M}(\mathbb{R}^2)$ is continuous. Since $\mathcal{L}^1_{\mathrm{loc}}[0, M)$ and $\mathcal{M}(\mathbb{R}^2)$ are Fréchet spaces and $\mathcal{C}[0, M)$ is dense in $\mathcal{L}^1_{\mathrm{loc}}[0, M)$ (with respect to convergence in the topology of $\mathcal{L}^1_{\mathrm{loc}}[0, M)$) there exists a unique (uniformly) continuous extension of the mapping $\ell \mapsto \ell\mu$ to $\mathcal{L}^1_{\mathrm{loc}}[0, M)$, and (4.34) automatically holds for this extension. □

4.3.3. *Solution to the abstract age equation.* In Proposition 4.16 below, we establish an explicit one-to-one correspondence between solutions $\{\zeta_t\}_{t\geq 0}$ to the abstract age equation for some $\gamma \in \tilde{\mathcal{M}}$ and solutions $\{\mu_t\}_{t\geq 0}$ to the simplified age equation for a related $\tilde{\gamma} \in \tilde{\mathcal{M}}$. In order to state this correspondence, given $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$, we define

$$\psi_\ell(x, t) \doteq \exp(r_\ell(x, t)) \tag{4.45}$$

for $(x, t) \in \mathbb{R}^2$, where

$$r_\ell(x, t) \doteq \begin{cases} -\displaystyle\int_{x-t}^{x} \ell(u)\,du, & \text{if } 0 \leq t \leq x < M, \\ -\displaystyle\int_{0}^{x} \ell(u)\,du, & \text{if } 0 \leq x \leq t, x < M, \\ 0, & \text{otherwise.} \end{cases} \tag{4.46}$$

Note that $\psi_{-\ell} = \psi_\ell^{-1}$ and r_ℓ and ψ_ℓ are continuous, locally bounded functions on $(-\infty, M) \times [0, \infty)$. Hence, for every $t \in [0, \infty)$ and measure $\chi \in \tilde{\mathcal{M}}$, the measure $\psi_\ell\chi$ defined by $\psi_\ell\chi(\varphi) \doteq \chi(\psi_\ell\varphi)$, and, likewise, the measure $\psi_{-\ell}\chi$ lie in $\tilde{\mathcal{M}}$. Also, if ℓ is continuous, then $(\psi_\ell)_x + (\psi_\ell)_s$ exists and is continuous and satisfies

$$(\psi_\ell)_x + (\psi_\ell)_s = -\ell\psi_\ell. \tag{4.47}$$

PROPOSITION 4.16. *Given $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$ and $\{\zeta_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0, \infty)$, suppose that the measure $\ell\zeta$ defined by*

$$(\ell\zeta)(\varphi) \doteq \int_0^\infty \langle \ell(\cdot)\varphi(\cdot, s), \zeta_s\rangle\,ds, \qquad \varphi \in \tilde{\mathcal{C}}, \tag{4.48}$$

lies in $\tilde{\mathcal{M}}$. Then $\{\zeta_t\}_{t\geq 0}$ solves the abstract age equation for ℓ and $\gamma \in \tilde{\mathcal{M}}$ if and only if $\{\mu_t\}_{t\geq 0}$ defined by

$$\langle f, \mu_t\rangle \doteq \langle f(\cdot)\psi_{-\ell}(\cdot, t), \zeta_t\rangle \tag{4.49}$$

satisfies the simplified age equation for $\tilde{\gamma} \doteq \psi_{-\ell}\gamma$, where $(\psi_{-\ell}\gamma)(\varphi) \doteq \gamma(\psi_{-\ell}\varphi)$ for $\varphi \in \tilde{\mathcal{C}}$.

PROOF. Let $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$ and $\{\zeta_t\}_{t\geq 0} \in \mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$ be such that $\ell\zeta \in \tilde{\mathcal{M}}$. Moreover, assume that $\{\zeta_t\}_{t\geq 0}$ solves the abstract age equation for ℓ and $\gamma \in \tilde{\mathcal{M}}$. Since $\psi_{-\ell}$ is continuous, the function $\{\mu_t\}_{t\geq 0}$ defined in (4.49) lies in $\mathcal{D}_{\mathcal{M}[0,M)}[0,\infty)$. Choose a sequence of continuous functions $\ell_n, n \in \mathbb{N}$, defined on $[0, M)$ such that as $n \to \infty$, ℓ_n converges to ℓ in $\mathcal{L}^1_{\mathrm{loc}}[0, M)$. From (4.47) it follows that

$$(\psi_{-\ell_n})_x + (\psi_{-\ell_n})_s = \ell_n \psi_{-\ell_n}. \tag{4.50}$$

Given $\tilde{\varphi} \in \tilde{\mathcal{C}}^{1,1}$, let $\varphi \doteq \psi_{-\ell_n}\tilde{\varphi}$. Then φ clearly lies in $\tilde{\mathcal{C}}$. Moreover, because $\tilde{\varphi} \in \tilde{\mathcal{C}}^{1,1}$, ℓ_n is continuous and (4.50) is satisfied, it follows that $\varphi_x + \varphi_s$ exists and is continuous and, hence, that $\varphi \in \tilde{\mathcal{C}}^{1,1}$. Therefore, substituting $\varphi = \psi_{-\ell_n}\tilde{\varphi}$ into the abstract age equation (4.17) and using (4.50), we obtain

$$\begin{aligned} &-\int_0^\infty \langle \ell_n(\cdot)\psi_{-\ell_n}(\cdot,s)\tilde{\varphi}(\cdot,s), \zeta_s\rangle\, ds - \int_0^\infty \big\langle \psi_{-\ell_n}(\cdot,s)\big(\tilde{\varphi}_x(\cdot,s) + \tilde{\varphi}_s(\cdot,s)\big), \zeta_s\big\rangle ds \\ &\quad = -\int_0^\infty \langle \ell(\cdot)\psi_{-\ell_n}(\cdot,s)\tilde{\varphi}(\cdot,s), \zeta_s\rangle + \gamma(\tilde{\varphi}\psi_{-\ell_n}). \end{aligned}$$

Rewriting the right-hand side of the last equation using (4.49), we see that

$$\begin{aligned} &-\int_0^\infty \big\langle \psi_{\ell-\ell_n}(\cdot,s)\big(\tilde{\varphi}_x(\cdot,s) + \tilde{\varphi}_s(\cdot,s)\big), \mu_s\big\rangle ds \\ &\quad = \int_0^\infty \big\langle \big(\ell_n(\cdot) - \ell(\cdot)\big)\psi_{-\ell_n}(\cdot,s)\tilde{\varphi}(\cdot,s), \zeta_s\big\rangle ds + \tilde{\gamma}(\tilde{\varphi}\psi_{\ell-\ell_n}). \end{aligned} \tag{4.51}$$

As $n \to \infty$, $\psi_{\ell-\ell_n} \to 1$ uniformly on compact sets. As a result, we have

$$\begin{aligned} &\lim_{n\to\infty} \int_0^\infty \big\langle \psi_{\ell-\ell_n}(\cdot,s)\big(\tilde{\varphi}_x(\cdot,s) + \tilde{\varphi}_s(\cdot,s)\big), \mu_s\big\rangle ds \\ &\quad = \int_0^\infty \big\langle \big(\tilde{\varphi}_x(\cdot,s) + \tilde{\varphi}_s(\cdot,s)\big), \mu_s\big\rangle ds, \end{aligned} \tag{4.52}$$

and, due to the dominated convergence theorem, we have

$$\lim_{n\to\infty} \tilde{\gamma}(\tilde{\varphi}\psi_{\ell-\ell_n}) = \tilde{\gamma}(\tilde{\varphi}). \tag{4.53}$$

Furthermore, due to the assumption that $\ell\zeta \in \tilde{\mathcal{M}}$, Lemma 4.12 shows that $\{\zeta_t\}_{t\geq 0}$ satisfies the simplified age equation for $\ell\zeta + \gamma \in \tilde{\mathcal{M}}$. Proposition 4.15 can then be applied to conclude that for $(m_1, m_2) \in [0, M) \times [0,\infty)$, there exist $\tilde{L}(m_1, m_2) < \infty$ such that for every $\tilde{\varphi}$ with $\mathrm{supp}(\tilde{\varphi}) \subseteq \mathcal{R}_{m_1,m_2}$,

$$\begin{aligned} &\left|\int_0^\infty \big\langle \big(\ell(\cdot) - \ell_n(\cdot)\big)\psi_{-\ell_n}(\cdot,s)\tilde{\varphi}(\cdot,s), \zeta_s\big\rangle ds\right| \\ &\quad \leq \|\tilde{\varphi}\psi_{-\ell_n}\|_\infty \tilde{L}(m_1,m_2)\left(\int_{[0,M)} |\ell(x) - \ell_n(x)|\, dx\right). \end{aligned}$$

Due to the convergence of ℓ_n to ℓ in $\mathcal{L}^1_{\mathrm{loc}}[0, M)$ and the fact that $\|\tilde{\varphi}\psi_{-\ell_n}\|_\infty \to \|\tilde{\varphi}\psi_{-\ell}\|_\infty < \infty$, the right-hand side (and therefore the left-hand side) vanishes as $n \to \infty$. Taking limits as $n \to \infty$ in (4.51), the last assertion, together with (4.52) and (4.53), imply that $\{\mu_t\}_{t\geq 0}$ satisfies the simplified age equation for $\tilde{\gamma}$.

The converse is established in an exactly analogous fashion, and so we provide only a rough sketch of the proof. Suppose $\{\mu_t\}_{t\geq 0}$ satisfies the simplified age equation for $\tilde{\gamma}$. Then, given $\varphi \in \tilde{\mathcal{C}}^{1,1}$, substituting $\tilde{\varphi} \doteq \psi_{\ell_n}\varphi \in \tilde{\mathcal{C}}^{1,1}$ into the simplified age equation, using the PDE (4.47), (4.49) and the estimate from Proposition 4.15 and then sending $n \to \infty$, it can be shown that $\{\zeta_t\}_{t\geq 0}$ solves the abstract age equation for ℓ and γ. □

Combining Lemma 4.13 with Proposition 4.16, we then obtain the following characterization of solutions to the abstract age equation.

COROLLARY 4.17. *Given $\{\zeta_t\}_{t\geq 0}$ such that $\ell\zeta$ defined by (4.48) lies in $\tilde{\mathcal{M}}$, $\{\zeta_t\}_{t\geq 0}$ satisfies the abstract age equation for some $\gamma \in \tilde{\mathcal{M}}$ if and only if for every $t \in [0, \infty)$,*

$$\langle f, \zeta_t\rangle = \langle f(\cdot)\psi_\ell(\cdot, t), \mu_t\rangle = \gamma\big(\psi_{-\ell}\Lambda^t_{f(\cdot)\psi_\ell(\cdot,t)}\big). \tag{4.54}$$

4.3.4. *Proof of Theorem 4.1.* To begin with, note that by substituting $\ell = h$ into the definition (4.45) of ψ_ℓ, elementary calculations show that

$$\psi_h(x,t) = \begin{cases} \dfrac{1-G(x)}{1-G(x-t)}, & \text{if } 0 \leq t \leq x < M, \\ 1-G(x), & \text{if } 0 \leq x \leq t < \infty, \\ 0, & \text{otherwise.} \end{cases} \tag{4.55}$$

In particular, this implies that

$$\psi_h^{-1}(0,t) = \psi_h^{-1}(x,0) = 1, \qquad x, t \in [0, M) \times \mathbb{R}_+. \tag{4.56}$$

Now, assume that $\{\overline{\nu}_s\}_{s\geq 0}$ satisfies the condition (4.1). Then Lemma 4.12 shows that $\{\overline{\nu}_s\}_{s\geq 0}$ satisfies the age equation (4.2) for υ_0 and Z if and only if $\{\overline{\nu}_s\}_{s\geq 0}$ satisfies the abstract age equation for $\xi = \xi(\upsilon_0, Z)$ defined in (4.19). In turn, by Corollary 4.17 the latter statement holds if and only if

$$\langle f, \overline{\nu}_t\rangle = \xi\big(\psi_{-h}\Lambda^t_{f(\cdot)\psi_h(\cdot,t)}\big), \qquad t \geq 0,\ f \in \mathcal{C}_c[0, M). \tag{4.57}$$

However, for $x \in [0, M)$,

$$\big(\psi_{-h}\Lambda^t_{f(\cdot)\psi_h(\cdot,t)}\big)(x,0) = \psi_{-h}(x,0)f(x+t)\psi_h(x+t,t) = f(x+t)\frac{1-G(x+t)}{1-G(x)},$$

and for all $s \in [0, \infty)$,

$$\begin{aligned}\big(\psi_{-h}\Lambda^t_{f(\cdot)\psi h(\cdot,t)}\big)(0,s) &= \psi_{-h}(0,s)f(t-s)\psi_h(t-s,t)\mathbb{1}_{[0,t]}(s) \\ &= f(t-s)\big(1-G(t-s)\big)\mathbb{1}_{[0,t]}(s).\end{aligned}$$

Substituting this back into (4.57) and using the definition (4.19) of ξ, it follows that (4.57) coincides with the representation (4.3) for $\overline{v}_t$. This completes the proof of Theorem 4.1.

5. Functional law of large numbers limit. The main objective of this section is to show that, under suitable assumptions, the sequence $\{(\overline{X}^{(N)}, \overline{v}^{(N)})\}$ converges to a process that solves the fluid equations. In particular, this establishes existence of solutions to the fluid equations. First, in Section 5.1 we provide a useful description of the evolution of the state $(\overline{X}^{(N)}, \overline{v}^{(N)})$ of the N-server model. Then, in Section 5.2, we introduce a family of martingales that are used in Section 5.3 to establish tightness of the sequence $\{(\nu^{(N)}, X^{(N)})\}$. Finally, in Section 5.4, we provide the proof of Theorem 3.7.

5.1. *A characterization of the pre-limit processes.* The dynamics of the N-server model was described in Section 2.1 and certain auxiliary processes were introduced in Section 2.2. In this section, we provide a more succinct and convenient description of the state dynamics, which takes a form similar to that of the fluid equations.

Fix $N \in \mathbb{N}$ and, throughout the section, suppose $R_E^{(N)}$ and initial conditions $X^{(N)}(0) \in \mathbb{R}_+$ and $\nu_0^{(N)} \in \mathcal{M}_{\le N}[0, M)$ are given, and let $E^{(N)}$, $X^{(N)}$ and $\nu^{(N)}$ be the associated state processes, as described in Section 2.1. Recall that by the definition (2.7) of the age process, a customer j completes service (and therefore departs the system) at time s if and only if, at time s, the left derivative of the age process $a_j^{(N)}$ is positive and the right derivative is zero. For any measurable function φ on $[0, M) \times \mathbb{R}_+$, consider the sequence of real-valued processes $\{Q_\varphi^{(N)}\}$ given by

$$
\begin{aligned}
Q_\varphi^{(N)}(t) \doteq \sum_{s\in[0,t]} \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)}\rangle+1}^{K^{(N)}(t)} & \mathbb{1}_{\{d/dt a_j^{(N)}(s-)>0, d/dt a_j^{(N)}(s+)=0\}} \\
& \times \varphi\big(a_j^{(N)}(s), s\big),
\end{aligned} \tag{5.1}
$$

where $K^{(N)}$ and $a_j^{(N)}$ are defined by the relations (2.6) and (2.7). It follows immediately from (5.1) and the right-continuity of the filtration $\{\mathcal{F}_t^{(N)}\}$ that $Q_\varphi^{(N)}$ is $\{\mathcal{F}_t^{(N)}\}$-adapted. Also, from relations (2.6)–(2.8), it is easy to see that $Q_{\mathbf{1}}^{(N)}$ is equal to the cumulative departure process $D^{(N)}$ defined in (2.5) and that for every $N \in \mathbb{N}$, bounded, measurable φ and $t \in [0, \infty)$,

$$
\begin{aligned}
\big|Q_\varphi^{(N)}(t)\big| &\le \|\varphi\|_\infty \big(\langle \mathbf{1}, \nu_0^{(N)}\rangle + K^{(N)}(t)\big) \\
&\le \|\varphi\|_\infty \big(X^{(N)}(0) + E^{(N)}(t)\big).
\end{aligned} \tag{5.2}
$$

Dividing (5.2) by N, taking first expectations and then the supremum over N, by Remark 3.1, we also have

$$\sup_N \mathbb{E}\big[|\overline{Q}_\varphi^{(N)}(t)|\big] \le \|\varphi\|_\infty \sup_N \big(\mathbb{E}\big[\overline{X}^{(N)}(0)\big] + \mathbb{E}\big[\overline{E}^{(N)}(t)\big]\big) < \infty. \tag{5.3}$$

We now state the main result of this section. Recall that for $r, s \in [0,\infty)$, $\nu_s^{(N)}$ represents $\nu^{(N)}(s)$ and $\langle \varphi(\cdot + r, s), \nu_s^{(N)}\rangle$ is used as a short form for $\int_{[0,M)} \varphi(x + r, s)\nu_s^{(N)}(dx)$.

THEOREM 5.1. *The processes $(E^{(N)}, X^{(N)}, \nu^{(N)})$ satisfy a.s. the following coupled set of equations: for $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$ and $t \in [0,\infty)$,*

$$\begin{aligned}\langle \varphi(\cdot, t), \nu_t^{(N)}\rangle &= \langle \varphi(\cdot, 0), \nu_0^{(N)}\rangle + \int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \nu_s^{(N)}\rangle ds \\ &\quad - Q_\varphi^{(N)}(t) + \int_{[0,t]} \varphi(0, u)\, dK^{(N)}(u),\end{aligned} \tag{5.4}$$

$$X^{(N)}(t) = X^{(N)}(0) + E^{(N)}(t) - Q_{\mathbf{1}}^{(N)}(t), \tag{5.5}$$

$$N - \langle 1, \nu_t^{(N)}\rangle = \big[N - X^{(N)}(t)\big]^+, \tag{5.6}$$

where $K^{(N)}$ satisfies (2.6), and $Q_\varphi^{(N)}$ is the process defined in (5.1).

The rest of this section is devoted to the proof of this theorem. Fix $\omega \in \Omega$ (we will later restrict ourselves to ω in a set of probability 1 on which Lemmas 5.2 and 5.3 apply). We start with the simple observation that for any $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$, due to the right-continuity of $\nu^{(N)}$ we have for any $t \in [0,\infty)$,

$$\begin{aligned}&\langle \varphi(\cdot, t), \nu_t^{(N)}\rangle - \langle \varphi(\cdot, 0), \nu_0^{(N)}\rangle \\ &\quad = \lim_{n\to\infty} \sum_{k=0}^{\lfloor nt\rfloor} \left[\left\langle \varphi\left(\cdot, \frac{k+1}{n}\right), \nu_{(k+1)/n}^{(N)}\right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_{k/n}^{(N)}\right\rangle\right].\end{aligned} \tag{5.7}$$

In order to compute the increments on the right-hand side of the last equation, we observe that for $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$, $n \in \mathbb{N}$ and $k = 0, \ldots, \lfloor nt\rfloor$, we can write

$$\begin{aligned}&\left\langle \varphi\left(\cdot, \frac{k+1}{n}\right), \nu_{(k+1)/n}^{(N)}\right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_{k/n}^{(N)}\right\rangle \\ &\quad = \left\langle \varphi\left(\cdot, \frac{k+1}{n}\right) - \varphi\left(\cdot, \frac{k}{n}\right), \nu_{(k+1)/n}^{(N)}\right\rangle \\ &\qquad + \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_{(k+1)/n}^{(N)}\right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_{k/n}^{(N)}\right\rangle.\end{aligned} \tag{5.8}$$

Summing the first term on the right-hand side of (5.8) over $k = 0, \dots, \lfloor nt \rfloor$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{\lfloor nt \rfloor} \left\langle \varphi\left(\cdot, \frac{k+1}{n}\right) - \varphi\left(\cdot, \frac{k}{n}\right), \nu^{(N)}_{(k+1)/n} \right\rangle \\
&\quad = \sum_{k=1}^{\lfloor nt \rfloor} \left\langle \varphi\left(\cdot, \frac{k}{n}\right) - \varphi\left(\cdot, \frac{k-1}{n}\right), \nu^{(N)}_{k/n} \right\rangle \\
&\qquad + \left\langle \varphi\left(\cdot, \frac{\lfloor nt \rfloor + 1}{n}\right) - \varphi\left(\cdot, \frac{\lfloor nt \rfloor}{n}\right), \nu^{(N)}_{(\lfloor nt \rfloor + 1)/n} \right\rangle.
\end{aligned} \tag{5.9}$$

In order to simplify the last two terms on the right-hand side of (5.8), we first observe that for $\varphi \in \mathcal{C}^{1,1}_c([0, M) \times \mathbb{R}_+)$, $\delta \in (0, M)$ and $s \in [0, \infty)$, we have

$$\langle \varphi(\cdot, s), \nu^{(N)}_{s+\delta} \rangle = \mathcal{I}_1 + \mathcal{I}_2, \tag{5.10}$$

where

$$\mathcal{I}_1 \doteq \int_{[\delta, M)} \varphi(x, s) \nu^{(N)}_{s+\delta}(dx) \quad \text{and} \quad \mathcal{I}_2 \doteq \int_{[0,\delta)} \varphi(x, s) \nu^{(N)}_{s+\delta}(dx).$$

We begin by rewriting $\mathcal{I}_1$ in terms of quantities that are known at time s. For $x \geq \delta$, customers in service with age equal to x at time $s + \delta$ are precisely those customers that were already in service at time s with age equal to $x - \delta \geq 0$ and that, in addition, did not depart the system in the interval $[s, s + \delta]$. From (2.7) it is clear that the age of a customer already in service increases linearly with rate 1. Therefore, using the representation for $\nu^{(N)}$ given in (2.8), we have

$$\begin{aligned}
\mathcal{I}_1 &= \sum_{j=-\langle \mathbf{1}, \nu^{(N)}_0 \rangle + 1}^{K^{(N)}(s+\delta)} \varphi\big(a^{(N)}_j(s+\delta), s\big) \mathbb{1}_{\{\delta \leq a^{(N)}_j(s+\delta) < v_j\}} \\
&= \sum_{j=-\langle \mathbf{1}, \nu^{(N)}_0 \rangle + 1}^{K^{(N)}(s)} \varphi\big(a^{(N)}_j(s)+\delta, s\big) \mathbb{1}_{\{a^{(N)}_j(s)+\delta < v_j\}} \\
&= \sum_{j=-\langle \mathbf{1}, \nu^{(N)}_0 \rangle + 1}^{K^{(N)}(s)} \varphi\big(a^{(N)}_j(s)+\delta, s\big) \mathbb{1}_{\{a^{(N)}_j(s) < v_j\}} \\
&\quad - \sum_{j=-\langle \mathbf{1}, \nu^{(N)}_0 \rangle + 1}^{K^{(N)}(s)} \varphi\big(a^{(N)}_j(s)+\delta, s\big) \mathbb{1}_{\{a^{(N)}_j(s) < v_j \leq a^{(N)}_j(s)+\delta\}}.
\end{aligned}$$

(Here, and in what follows below, we always assume, without loss of generality, that $\delta = \delta(\omega)$ is sufficiently small so that the range of the first argument of φ

falls within $[0, M)$, ensuring that all quantities are well defined.) Substituting the definition of $\nu_s^{(N)}$ into the last expression, this can be rewritten as

$$\mathcal{I}_1 = \langle \varphi(\cdot + \delta, s), \nu_s^{(N)} \rangle - q_\varphi^{(N)}(s, \delta), \tag{5.11}$$

where for $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$, $s \in [0, \infty)$ and $\delta > 0$, we define

$$q_\varphi^{(N)}(s, \delta) \doteq \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{K^{(N)}(s)} \varphi\big(a_j^{(N)}(s) + \delta, s\big) \mathbb{1}_{\{a_j^{(N)}(s) < v_j \le a_j^{(N)}(s) + \delta\}}. \tag{5.12}$$

We now expand $\mathcal{I}_2$ as follows:

$$\mathcal{I}_2 = \int_{[0,\delta)} \varphi(0, s) \nu_{s+\delta}^{(N)}(dx) + \int_{[0,\delta)} \big(\varphi(x, s) - \varphi(0, s)\big) \nu_{s+\delta}^{(N)}(dx).$$

For any $0 \le r \le s < \infty$, let $K^{(N)}(r, s] = K^{(N)}(s) - K^{(N)}(r)$ denote the number of customers that entered service in the period $(r, s]$, and let $D_*^{(N)}(r, s]$ denote the number of customers that both entered service and departed the system in the period $(r, s]$. Note that $D_*^{(N)}(r, s]$ admits the explicit representation

$$D_*^{(N)}(r, s] = \sum_{j=K^{(N)}(r)+1}^{K^{(N)}(s)} \mathbb{1}_{\{a_j^{(N)}(s) = v_j\}}. \tag{5.13}$$

Also, note that $\nu_{s+\delta}^{(N)}[0, \delta)$ is the number of customers in service that have age less than δ at time $s + \delta$. These customers must therefore have entered service in the interval $(s, s + \delta]$ and not yet departed by time $s + \delta$. Therefore, we can write

$$\nu_{s+\delta}^{(N)}[0, \delta) = K^{(N)}(s, s + \delta] - D_*^{(N)}(s, s + \delta].$$

Combining the last three expressions, we obtain

$$\begin{aligned} \mathcal{I}_2 = {} & \varphi(0, s) K^{(N)}(s, s + \delta] - \varphi(0, s) D_*^{(N)}(s, s + \delta] \\ & + \int_{[0,\delta]} \big(\varphi(x, s) - \varphi(0, s)\big) \nu_{s+\delta}^{(N)}(dx). \end{aligned} \tag{5.14}$$

Substituting (5.11) and (5.14) into (5.10), with $s = k/n$ and $\delta = 1/n$, we obtain for $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$,

$$\begin{aligned} & \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_{(k+1)/n}^{(N)} \right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu_{k/n}^{(N)} \right\rangle \\ & \quad = \left\langle \varphi\left(\cdot + \frac{1}{n}, \frac{k}{n}\right) - \varphi\left(\cdot, \frac{k}{n}\right), \nu_{k/n}^{(N)} \right\rangle - q_\varphi\left(\frac{k}{n}, \frac{1}{n}\right) \\ & \qquad + \varphi\left(0, \frac{k}{n}\right) K^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] - \varphi\left(0, \frac{k}{n}\right) D_*^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] \\ & \qquad + \int_{[0,1/n)} \left(\varphi\left(x, \frac{k}{n}\right) - \varphi\left(0, \frac{k}{n}\right)\right) \nu_{(k+1)/n}^{(N)}(dx). \end{aligned}$$

Summing the last expression over $k = 0, \dots, \lfloor nt \rfloor$, we obtain

$$\begin{aligned}
&\sum_{k=0}^{\lfloor nt \rfloor} \left[\left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu^{(N)}_{(k+1)/n} \right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu^{(N)}_{k/n} \right\rangle \right] \\
&\quad = \sum_{k=0}^{\lfloor nt \rfloor} \frac{1}{n} \left\langle \frac{\varphi(\cdot + 1/n, k/n) - \varphi(\cdot, k/n)}{1/n}, \nu^{(N)}_{k/n} \right\rangle - Q^{(N),1/n}_{\varphi}(t) \\
&\qquad + \sum_{k=0}^{\lfloor nt \rfloor} \varphi\left(0, \frac{k}{n}\right) K^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] - R^{(N),n}_{\varphi}(t),
\end{aligned} \tag{5.15}$$

where, for conciseness, we set, for $\varphi \in \mathcal{C}^{1,1}_c([0, M) \times \mathbb{R}_+)$, $N \in \mathbb{N}$ and $t \in [0, \infty)$,

$$Q^{(N),\delta}_{\varphi}(t) \doteq \sum_{k=0}^{\lfloor t/\delta \rfloor} q^{(N)}_{\varphi}(k\delta, \delta), \qquad \delta > 0, \tag{5.16}$$

and, for $n \in \mathbb{N}$,

$$\begin{aligned}
R^{(N),n}_{\varphi}(t) &\doteq \sum_{k=0}^{\lfloor nt \rfloor} \varphi\left(0, \frac{k}{n}\right) D^{(N)}_{*}\left(\frac{k}{n}, \frac{k+1}{n}\right] \\
&\quad - \sum_{k=0}^{\lfloor nt \rfloor} \int_{[0,1/n)} \left(\varphi\left(x, \frac{k}{n}\right) - \varphi\left(0, \frac{k}{n}\right) \right) \nu^{(N)}_{(k+1)/n}(dx).
\end{aligned} \tag{5.17}$$

Summing (5.8) over $k = 0, 1, \dots, \lfloor nt \rfloor$, and using (5.9) and (5.15), we obtain

$$\begin{aligned}
&\sum_{k=0}^{\lfloor nt \rfloor} \left[\left\langle \varphi\left(\cdot, \frac{k+1}{n}\right), \nu^{(N)}_{(k+1)/n} \right\rangle - \left\langle \varphi\left(\cdot, \frac{k}{n}\right), \nu^{(N)}_{k/n} \right\rangle \right] \\
&\quad = \left\langle \varphi\left(\cdot + \frac{1}{n}, 0\right) - \varphi(\cdot, 0), \nu^{(N)}_0 \right\rangle \\
&\qquad + \left\langle \varphi\left(\cdot, \frac{\lfloor nt \rfloor + 1}{n}\right) - \varphi\left(\cdot, \frac{\lfloor nt \rfloor}{n}\right), \nu^{(N)}_{(\lfloor nt \rfloor + 1)/n} \right\rangle \\
&\qquad + \sum_{k=1}^{\lfloor nt \rfloor} \left\langle \varphi\left(\cdot + \frac{1}{n}, \frac{k}{n}\right) - \varphi\left(\cdot, \frac{k-1}{n}\right), \nu^{(N)}_{k/n} \right\rangle \\
&\qquad + \sum_{k=0}^{\lfloor nt \rfloor} \varphi\left(0, \frac{k}{n}\right) K^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] \\
&\qquad - Q^{(N),1/n}_{\varphi}(t) - R^{(N),n}_{\varphi}(t).
\end{aligned} \tag{5.18}$$

Because $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$ and $\nu^{(N)}[0, M) \le N$, by the bounded convergence theorem, it follows that (for the fixed ω)

$$\begin{aligned} \lim_{n\to\infty} \left\langle \varphi\left(\cdot + \frac{1}{n}, 0\right) - \varphi(\cdot, 0), \nu_0^{(N)} \right\rangle = 0, \\ \lim_{n\to\infty} \left\langle \varphi\left(\cdot, \frac{\lfloor nt \rfloor + 1}{n}\right) - \varphi\left(\cdot, \frac{\lfloor nt \rfloor}{n}\right), \nu^{(N)}_{(\lfloor nt \rfloor + 1)/n} \right\rangle = 0. \end{aligned} \tag{5.19}$$

Therefore, the first two terms on the right-hand side of (5.18) vanish as $n \to \infty$. Next, multiplying and dividing the third and fourth terms on the right-hand side of (5.18) by $1/n$, and taking limits as $n \to \infty$, we obtain the corresponding Riemann–Stieltjes integrals

$$\begin{aligned} &\lim_{n\to\infty} \sum_{k=1}^{\lfloor nt \rfloor} \frac{1}{n} \left\langle \frac{\varphi(\cdot + 1/n, k/n) - \varphi(\cdot, (k-1)/n)}{1/n}, \nu_{k/n}^{(N)} \right\rangle \\ &\quad = \int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \nu_s^{(N)} \rangle \, ds \end{aligned} \tag{5.20}$$

and

$$\begin{aligned} &\lim_{n\to\infty} \sum_{k=0}^{\lfloor nt \rfloor} \varphi\left(0, \frac{k}{n}\right) K^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] \\ &\quad = \int_{[0,t]} \varphi(0, s) \, dK^{(N)}(s), \end{aligned} \tag{5.21}$$

where we have used the fact that the process $K^{(N)}$ has right-continuous paths in the latter limit. The next two results identify the limits, as $n \to \infty$, of the remaining two terms, $Q_\varphi^{(N),1/n}$ and $R^{(N),n}$, on the right-hand side of (5.18).

LEMMA 5.2. *Almost surely, for every* $N \in \mathbb{N}$, $t \in [0, \infty)$ *and* $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$, $Q_\varphi^{(N),\delta}(t)$ *converges to* $Q_\varphi^{(N)}(t)$ *as* $\delta \to 0$.

PROOF. Fix $N \in \mathbb{N}$, $t \in [0, \infty)$ and $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$, and let $L < \infty$ be such that $\sup_{s\in[0,t], y\in[0,M)} |\varphi_x(y, s) + \varphi_t(y, s)| \le L$. For any $\delta > 0$, and $j = -\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1, \ldots, 0$, define $\tau^\delta(j) = 0$ and for $j = 1, 2, \ldots$, define

$$\tau^\delta(j) \doteq \inf\{k \in \mathbb{N} : a_j^{(N)}(k\delta + \varepsilon) > 0 \ \forall \varepsilon > 0\}.$$

Observe that $\tau^\delta(j)\delta$ represents the smallest point on the δ-lattice $\{k\delta, k = 0, 1, \ldots, \lfloor t/\delta \rfloor\}$ that is greater than or equal to the time at which the jth customer enters service. The introduction of ε in the definition of $\tau^\delta(j)$ was necessary to ensure that $\tau^\delta(j) = k$ if the jth customer enters service precisely at $k\delta$, and thus

has age 0 at that time. For any $\delta > 0$, a simple interchange of summation shows that

$$Q_\varphi^{(N),\delta}(t) = \sum_{k=0}^{\lfloor t/\delta \rfloor} \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{K^{(N)}(k\delta)} \varphi\big(a_j^{(N)}(k\delta) + \delta, k\delta\big) \mathbb{1}_{\{a_j^{(N)}(k\delta) < v_j \le a_j^{(N)}(k\delta)+\delta\}}$$

$$= \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{K^{(N)}(\lfloor t/\delta \rfloor \delta)} \sum_{k=\tau^\delta(j)}^{\lfloor t/\delta \rfloor} \varphi\big(a_j^{(N)}(k\delta) + \delta, k\delta\big) \mathbb{1}_{\{a_j^{(N)}(k\delta) < v_j \le a_j^{(N)}(k\delta)+\delta\}}.$$

However, when $a_j^{(N)}(k\delta) < v_j \le a_j^{(N)}(k\delta) + \delta$, we have

$$\sup_{s \in [k\delta, (k+1)\delta]} \big|\varphi\big(a_j^{(N)}(k\delta) + \delta, k\delta\big) - \varphi(v_j, s)\big| \le L\delta,$$

and we also know that there exists a (unique) $s \in (k\delta, (k+1)\delta]$ such that $\frac{d}{dt} a_j^{(N)}(s-) > 0$ and $a_j^{(N)}(s) = v_j$ (i.e., s is the unique time at which the customer departs the system). Hence, we can write

$$Q_\varphi^{(N),\delta}(t) = \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{K^{(N)}(\lfloor t/\delta \rfloor \delta)} \sum_{s \in [0, (\lfloor t/\delta \rfloor + 1)\delta]} \varphi(v_j, s) \mathbb{1}_{\{d/dt a_j^{(N)}(s-) > 0, a_j^{(N)}(s) = v_j\}} + O(\delta).$$

Sending $\delta \to 0$, because $K^{(N)}$ is càdlàg, we see that $Q_\varphi^{(N),\delta}(t)$ converges to the quantity

$$\sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{K^{(N)}(t-)} \sum_{s \in [0,t]} \varphi\big(a_j^{(N)}(s), s\big) \mathbb{1}_{\{d/dt a_j^{(N)}(s-) > 0, a_j^{(N)}(s) = v_j\}} = Q_\varphi^{(N)}(t),$$

where the last equality follows by replacing $K^{(N)}(t-)$ by $K^{(N)}(t)$. This replacement is justified even though we need not have $K^{(N)}(t-) = K^{(N)}(t)$ because the relation $G(0+) = 0$ ensures that every v_j is almost surely strictly positive, and so if customer j enters service precisely at time t, then $\mathbb{1}_{\{a_j^{(N)}(s) = v_j\}} = 0$ for every $s \in [0, t]$. $\square$

LEMMA 5.3. *Almost surely, for every $T \in [0, \infty)$ and every $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$,*

$$\lim_{n \to \infty} \sup_{t \in [0,T]} R_\varphi^{(N),n}(t) = 0. \tag{5.22}$$

PROOF. We will establish the lemma by showing that a.s. for any $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$, as $n \to \infty$, both terms on the right-hand side of (5.17) con-

verge uniformly to zero. Fix $T < \infty$ and $t \in [0, T]$. Then for any $n \in \mathbb{N}$, from the representation (5.13) for $D_*^{(N)}$ it immediately follows that

$$\sum_{k=0}^{\lfloor nt \rfloor} D_*^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] \le \sum_{k=0}^{\lfloor nt \rfloor} \sum_{j=K^{(N)}(k/n)+1}^{K^{(N)}((k+1)/n)} \mathbb{1}_{\{v_j \le 1/n\}} = \sum_{j=1}^{K^{(N)}((\lfloor nt \rfloor+1)/n)} \mathbb{1}_{\{v_j \le 1/n\}}.$$

We take first the supremum over $t \in [0, T]$ and then the expectation of both sides above. Using the independence of the interarrival and service times and the fact that $K^{(N)} \le X^{(N)}(0) + E^{(N)}$, which can be deduced from the relations (2.5), (2.6) and the fact that (2.10) implies $\langle \mathbf{1}, \nu^{(N)} \rangle \le X^{(N)}$, we then obtain the bound

$$\mathbb{E}\left[\sup_{t\in[0,T]} \sum_{k=0}^{\lfloor nt \rfloor} D_*^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right]\right] \le G\left(\frac{1}{n}\right)\mathbb{E}\big[X^{(N)}(0) + E^{(N)}(T+1)\big].$$

Next, taking the limit as $n \to \infty$, observe that the right-hand side tends to zero because $\mathbb{E}[X^{(N)}(0) + E^{(N)}(T+1)] < \infty$, G is right-continuous and $G(0+) = 0$. At the same time, on the left-hand side, the expectation can be interchanged with the limit by an application of the dominated convergence theorem because $D_*^{(N)}(k/n, (k+1)/n]$ is nonnegative for all $k \in \mathbb{N}$, and is bounded above by $K^{(N)}(T) + 1$, which has finite expectation. Thus,

$$\begin{aligned} 0 &= \lim_{n\to\infty} \mathbb{E}\left[\sup_{t\in[0,T]} \sum_{k=0}^{\lfloor nt \rfloor} D_*^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right]\right] \\ &= \mathbb{E}\left[\lim_{n\to\infty} \sup_{t\in[0,T]} \sum_{k=0}^{\lfloor nt \rfloor} D_*^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right]\right]. \end{aligned}$$

Since each term of the form $D_*^{(N)}(\frac{k}{n}, \frac{k+1}{n}]$ is nonnegative, this implies that the limit within the expectation on the right-hand side is almost surely zero. Therefore, almost surely,

$$\begin{aligned} &\lim_{n\to\infty} \sup_{t\in[0,T]} \sum_{k=0}^{\lfloor nt \rfloor} \varphi\left(0, \frac{k}{n}\right) D_*^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] \\ &\quad \le \|\varphi\|_\infty \lim_{n\to\infty} \sup_{t\in[0,T]} \sum_{k=0}^{\lfloor nt \rfloor} D_*^{(N)}\left(\frac{k}{n}, \frac{k+1}{n}\right] = 0. \end{aligned}$$

The monotonicity in T of the left-hand side allows us to conclude that there exists a set Ω_1 of full $\mathbb{P}$-measure on which this convergence holds simultaneously for all T.

We now turn to the second term on the right-hand side of (5.17). Let $m(\delta) \doteq \sup_{(x,t),(y,s)\in[0,M)\times\mathbb{R}_+ : |(x,t)-(y,s)|\le\delta} |\varphi(x,t) - \varphi(y,s)|$ be the modulus of continuity of φ. Note that $\lim_{\delta\to 0} m(\delta) = 0$ because $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$ is uniformly

continuous. For any $t \in [0, T]$ and $n \in \mathbb{N}$,

$$\sum_{k=0}^{\lfloor nt \rfloor} \int_{[0,1/n)} \left| \varphi\left(x, \frac{k}{n}\right) - \varphi\left(0, \frac{k}{n}\right) \right| \nu^{(N)}_{(k+1)/n}(dx) \le m\left(\frac{1}{n}\right) \sum_{k=0}^{\lfloor nt \rfloor} \nu^{(N)}_{(k+1)/n}\left[0, \frac{1}{n}\right).$$

Now, any customer whose age lies in $[0, 1/n)$ at time $(k+1)/n$ entered service strictly after k/n and would have age greater than or equal to $1/n$ at any time k'/n, $k+1 < k' \in \mathbb{N}$, if it were still in service at that time. Hence, for any fixed $n \in \mathbb{N}$, the unit mass corresponding to any given customer is counted in at most one term of the form $\nu^{(N)}_{(k+1)/n}[0, 1/n)$, $k \in \mathbb{N}$. This implies the elementary bound

$$\sup_{t \in [0,T]} \sum_{k=0}^{\lfloor nt \rfloor} \nu^{(N)}_{(k+1)/n}\left[0, \frac{1}{n}\right) \le X^{(N)}(0) + E^{(N)}(T+1).$$

Let Ω_2 be the set of full $\mathbb{P}$-measure on which the property $X^{(N)}(0) + E^{(N)}(t) < \infty$ for every $t \in [0, \infty)$ is satisfied. Then on Ω_2, the right-hand side of the last expression, which is independent of n, is a.s. finite. Therefore, a.s.,

$$\lim_{n \to \infty} \sup_{t \in [0,T]} \left| \sum_{k=0}^{\lfloor nt \rfloor} \int_{[0,1/n)} \left| \varphi\left(x, \frac{k}{n}\right) - \varphi\left(0, \frac{k}{n}\right) \right| \nu^{(N)}_{(k+1)/n}(dx) \right|$$

$$\le \left(X^{(N)}(0) + E^{(N)}(T+1)\right) \lim_{n \to \infty} m\left(\frac{1}{n}\right) = 0.$$

Thus, we have shown that on the set $\Omega_1 \cap \Omega_2$ of full $\mathbb{P}$-measure, (5.22) holds for every $T < \infty$ and $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$. □

We are now in a position to complete the proof of Theorem 5.1.

PROOF OF THEOREM 5.1. Fix $N \in \mathbb{N}$ and $t \in [0, \infty)$, choose a set $\tilde{\Omega}$ of $\mathbb{P}$-measure 1 on which the assertions of Lemmas 5.2 and 5.3 hold for every $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$. Fix $\omega \in \tilde{\Omega}$. Combining (5.18) with (5.19), (5.20), (5.21), Lemmas 5.2 and 5.3, it follows that the right-hand side of (5.7) equals

$$\int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \nu^{(N)}_s \rangle \, ds - Q^{(N)}_\varphi(t) + \int_{[0,t]} \varphi(0, s) \, dK^{(N)}(s).$$

Equating this with the left-hand side of (5.7), we obtain (5.4). The remaining relations (5.5) and (5.6) follow immediately from (2.5), (2.10) and the observation that $Q^{(N)}_{\mathbf{1}} = D^{(N)}$ and the observation made in the dicussion below (5.1) that $Q^{(N)}_{\mathbf{1}} = D^{(N)}$. □

5.2. *A useful family of martingales.* An inspection of the integral equation (5.4) suggests that identification of the limit of the sequence $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)})\}$ of

scaled state processes is likely to require a characterization of the limit of a scaled version $\overline{Q}_\varphi^{(N)}$ of $Q_\varphi^{(N)}$. In order to achieve this task, we first identify the compensator of $Q_\varphi^{(N)}$ (in Corollary 5.5) and then identify the limit of the quadratic variation of the associated scaled martingale $\overline{M}_\varphi^{(N)}$, obtained as a compensated sum of jumps (see Lemma 5.9).

We begin by introducing some notation. Recall that h is the hazard rate function defined in (2.3). For any $\varphi \in \mathcal{C}_b([0, M) \times \mathbb{R}_+)$, consider the sequences of processes $\{A_\varphi^{(N)}\}$ defined by

$$A_\varphi^{(N)}(t) \doteq \int_0^t \left(\int_{[0,M)} \varphi(x, s) h(x) \nu_s^{(N)}(dx) \right) ds \tag{5.23}$$

for every $t \in [0, \infty)$. We now derive an alternative representation for the process $A_\varphi^{(N)}$ which shows, in particular, that $A_\varphi^{(N)}$ is well defined and takes values in $\mathbb{R}$ for every $t \in [0, \infty)$. For $j \in \mathbb{N}$, recall that $\alpha_j^{(N)} \doteq \mathrm{inv}[K^{(N)}](j)$ is the time that the jth customer entered service. Interchanging the order of integration and summation and using the linear increase of the age process, we can write for $t \in [0, \infty)$,

$$\begin{aligned} A_\varphi^{(N)}(t) &= \int_0^t \left(\sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{K^{(N)}(s)} h\big(a_j^{(N)}(s)\big) \varphi\big(a_j^{(N)}(s), s\big) \mathbb{1}_{\{a_j^{(N)}(s) < v_j\}} \right) ds \\ &= \sum_{j=-\langle \mathbf{1}, \nu_0^{(N)} \rangle + 1}^{0} \int_0^t h\big(a_j^{(N)}(0) + s\big) \varphi\big(a_j^{(N)}(0) + s, s\big) \mathbb{1}_{\{a_j^{(N)}(0) + s < v_j\}} \, ds \\ &\quad + \sum_{j=1}^{K^{(N)}(t)} \int_{\alpha_j^{(N)}}^t h\big(s - \alpha_j^{(N)}\big) \varphi\big(s - \alpha_j^{(N)}, s\big) \mathbb{1}_{\{s < \alpha_j^{(N)} + v_j\}} \, ds. \end{aligned} \tag{5.24}$$

This shows that $A_\varphi^{(N)}$ is well defined because $v_j < M$ a.s., and h is locally integrable on $[0, M)$. Moreover, using the inequality $\langle \mathbf{1}, \nu_0^{(N)} \rangle + K^{(N)}(t) \le X^{(N)}(0) + E^{(N)}(t)$, we have for every $N \in \mathbb{N}$ and $\varphi \in \mathcal{C}_c([0, M) \times \mathbb{R}_+)$ with $\mathrm{supp}(\varphi) \subset [0, m] \times \mathbb{R}_+$ for $m \in [0, M)$,

$$\big|A_\varphi^{(N)}(t)\big| \le \|\varphi\|_\infty \big(X^{(N)}(0) + E^{(N)}(t)\big) \left(\int_0^m h(x) \, dx \right), \qquad t \in [0, \infty), \tag{5.25}$$

which is finite due to the local integrability of h.

Now, let $J_t^{(N)}$ be the (random) set of jump points of the departure process $D^{(N)}$ up to time t

$$J_t^{(N)} \doteq \big\{ s \in [0, t] : D^{(N)}(s) \neq D^{(N)}(s-) \big\}$$

and set $J^{(N)} = J_\infty^{(N)} \doteq \bigcup_{t>0} J_t^{(N)}$. Recall that $Q_{\mathbf{1}}^{(N)} = D^{(N)}$. We start by identifying the (predictable) compensator of $D^{(N)}$ (see Section 3b of [10] for the definition).

LEMMA 5.4. *For every $N \in \mathbb{N}$, the process $A_{\mathbf{1}}^{(N)}$ is the $\{\mathcal{F}_t^{(N)}\}$-compensator of the departure process $D^{(N)}$. In other words, $A_{\mathbf{1}}^{(N)}$ is an increasing, $\{\mathcal{F}_t^{(N)}\}$-adapted process, with $\mathbb{E}[A_{\mathbf{1}}^{(N)}(t)] < \infty$ for every $t \in [0,\infty)$, such that for every nonnegative $\{\mathcal{F}_t^{(N)}\}$-predictable process H,*

$$\begin{aligned}\mathbb{E}\Big[\sum_{s\in J^{(N)}} H_s\Big] &= \mathbb{E}\Big[\int_0^\infty H_s\, dA_{\mathbf{1}}^{(N)}(s)\Big] \\ &= \mathbb{E}\Big[\int_0^\infty H_s\Big(\int_{[0,M)} h(x)\nu_s^{(N)}(dx)\Big)\, ds\Big].\end{aligned} \tag{5.26}$$

PROOF. Fix $N \in \mathbb{N}$, label the servers from $1,\dots,N$ and assume without loss of generality that, for $j = -\langle \mathbf{1}, \nu_0^{(N)}\rangle + 1, \dots, 0$, the jth customer in service at time 0 is being served at the k_jth station, where $k_j \doteq j + \langle \mathbf{1}, \nu_0^{(N)}\rangle$.

In order to prove the lemma, we shall find it convenient to introduce the following notation. For $k = 1,\dots,N$ and $n \in \mathbb{N}$, let $\theta_n^{(N),k}$ (resp., $\zeta_n^{(N),k}$) be the time at which the nth customer to be served at station k starts (resp., completes) service, where for $j = -\langle \mathbf{1}, \nu_0^{(N)}\rangle + 1, \dots, 0$, we set $\theta_1^{(N),k_j}$ equal to $-a_j^{(N)}(0)$. We also let $D^{(N),k}(t)$ represent the total number of customers that have departed from the kth station in the interval $[0,t]$. Then clearly $D^{(N),k}(0) = 0$ and

$$D^{(N)} = \sum_{k=1}^N D^{(N),k}.$$

For conciseness, for the rest of this proof we shall omit the explicit dependence of all quantities on N.

For $k = 1,\dots,N$, the process $D^k = D^{(N),k}$ admits the decomposition

$$D^k(t) = \sum_{n=1}^\infty [D^k(t\wedge\zeta_n^k) - D^k(t\wedge\theta_n^k)], \qquad t \in [0,\infty).$$

Define

$$D_n^k(t) \doteq D^k(t\wedge\zeta_n^k) - D^k(t\wedge\theta_n^k), \qquad t \in [0,\infty).$$

Observe that D_n^k is a point process with just one point representing the nth departure from station k. We claim (and justify below) that the $\{\mathcal{F}_t\}$-compensator of D_n^k is given by the process A_n^k that is defined, for $t \in [0,\infty)$, by

$$A_n^k(t) \doteq \begin{cases} 0, & \text{if } t \in [0, \theta_n^k \vee 0], \\ \displaystyle\int_{\theta_n^k\vee 0}^t h(u-\theta_n^k)\, du, & \text{if } t \in (\theta_n^k \vee 0, \zeta_n^k], \\ \displaystyle\int_{\theta_n^k\vee 0}^{\zeta_n^k} h(u-\theta_n^k)\, du, & \text{if } t \in (\zeta_n^k, \infty). \end{cases}$$

It is straightforward to verify that θ_n^k and ζ_n^k are both $\{\mathcal{F}_t\}$-stopping times (this can be done by rewriting the events $\{\theta_n^k \le t\}$ and $\{\zeta_n^k \le t\}$ as events involving $\{a_j^{(N)}(s), \Upsilon_j^{(N)}(s), s \in [0,t], j \in \{-N+1, \ldots, 0\} \cup \mathbb{N}\}$—the details are left to the reader). As a result, it follows that A_n^k is $\{\mathcal{F}_t\}$-adapted. Moreover, by definition, A_n^k is continuous and, hence, $\{\mathcal{F}_t\}$-predictable. Thus, to establish the claim that it is the $\{\mathcal{F}_t\}$-compensator for D_n^k, by Theorem 3.17 on page 32 of [10], it suffices to show that for every $\{\mathcal{F}_t\}$-stopping time T,

$$\mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T]}(s)\, dD_n^k(s)\right] = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T]}(s)\, dA_n^k(s)\right].$$

We will prove the result for the case when $\theta_n^k > 0$ (i.e., when $n \ge 2$ or $n = 1$ and $k \ne k_j$ for $j = -\langle \mathbf{1}, \nu_0^{(N)}\rangle + 1, \ldots, 0$). The result in the remaining cases also follows from the same argument, but with G replaced by the conditional distribution $[G(\cdot) - G(a_j^{(N)}(0))]/[1 - G(a_j^{(N)}(0))]\mathbb{1}_{[a_j^{(N)}(0),\infty)}$, which has hazard rate $h\mathbb{1}_{[a_j^{(N)}(0),\infty)}$. Fix n, k such that $\theta_n^k > 0$. Due to the monotone convergence theorem, it is clear that it in fact suffices to show that the above equality holds for any bounded $\{\mathcal{F}_t\}$-stopping time T. Now, note that, because neither D_n^k nor A_n^k increase outside $(\theta_n^k, \zeta_n^k]$, the last equation is equivalent to the relation

$$\mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T]\cap(\theta_n^k,\infty)}(s)\, dD_n^k(s)\right] = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T]\cap(\theta_n^k,\zeta_n^k]}(s)\, dA_n^k(s)\right].$$

However, the term on the left-hand side can be rewritten as

$$\mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T]\cap(\theta_n^k,\infty)}(s)\, dD_n^k(s)\right] = \lim_{m\to\infty} \mathbb{E}\left[\sum_{j=0}^{\infty} \mathbb{1}_{\{\theta_n^k \le j/2^m < T,\, j/2^m < \zeta_n^k \le (j+1)/2^m\}}\right].$$

Since T, θ_n^k and ζ_n^k are all $\{\mathcal{F}_t\}$-stopping times, conditioning on $\mathcal{F}_{j/2^m}$, it follows that for any $m \in \mathbb{N}$ and $j = 1, \ldots, 2^m$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{\{\theta_n^k \le j/2^m < T\}} \mathbb{1}_{\{j/2^m < \zeta_n^k \le (j+1)/2^m\}}\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\{\theta_n^k \le j/2^m < T,\, \zeta_n^k > j/2^m\}} \mathbb{1}_{\{\zeta_n^k \le (j+1)/2^m\}} \mid \mathcal{F}_{j/2^m}\right]\right] \\
&\quad = \mathbb{E}\left[\mathbb{1}_{\{\theta_n^k \le j/2^m < T,\, \zeta_n^k > j/2^m\}} \int_{j/2^m}^{(j+1)/2^m} \frac{g(u - \theta_n^k)}{1 - G(j/2^m - \theta_n^k)}\, du\right] \\
&\quad = \mathbb{E}\left[\mathbb{1}_{\{\theta_n^k \le j/2^m < T,\, \zeta_n^k > j/2^m\}} \frac{G((j+1)/2^m - \theta_n^k) - G(j/2^m - \theta_n^k)}{1 - G(j/2^m - \theta_n^k)}\right].
\end{aligned}$$

Note that the second equality above uses the independence of the service requirement of a given customer from the cumulative arrival process and the service requirements of all other customers. Combining the last two displays and invoking

the monotone convergence theorem to justify the interchange of expectation and summation over j, we conclude that

$$\mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T]\cap(\theta_n^k,\infty)}(s)\,dD_n^k(s)\right]$$
$$= \lim_{m\to\infty} \mathbb{E}\left[\sum_{j=0}^{\infty} \mathbb{1}_{\{\theta_n^k \le j/2^m < T, \xi_n^k > j/2^m\}} \frac{G((j+1)/2^m - \theta_n^k) - G(j/2^m - \theta_n^k)}{1 - G(j/2^m - \theta_n^k)}\right].$$

To complete the proof, it only remains to show that the right-hand side of the last equation is equal to $\mathbb{E}[\int_0^\infty \mathbb{1}_{[0,T)}(s)A_n^k(ds)]$. For this, first note that the term within the expectation on the right-hand side of the last equation can be rewritten in the form $\mathbb{E}[\int_0^\infty \mathbb{1}_{[0,T)}(s)A_{m,n}^k(ds)]$ where, for $m \in \mathbb{N}$, $A_{m,n}^k(\cdot) = A_{m,n}^k(\omega;\cdot)$ is the random measure defined for each $\omega \in \Omega$ by

$$A_{m,n}^k(\omega; ds) = \sum_{j=0}^{\infty} \delta_{j/2^m}(ds)\mathbb{1}_{\{\theta_n^k(\omega) \le j/2^m < \zeta_n^k(\omega)\}}$$
$$\times \frac{G((j+1)/2^m - \theta_n^k(\omega)) - G(j/2^m - \theta_n^k(\omega))}{1 - G(j/2^m - \theta_n^k(\omega))},$$

where δ_x is, as usual, the Dirac mass at x. Next, observe that if $\theta_n^k + 2^{-m} \le u \le j2^{-m}$, then $G(u - \theta_n^k) \le G(j2^{-m} - \theta_n^k)$ and therefore $1 - G(u - \theta_n^k) \ge 1 - G(j2^{-m} - \theta_n^k)$ and similarly, if $\theta_n^k \le j2^{-m} \le u$ then $1 - G(u - \theta_n^k) \le 1 - G(j2^{-m} - \theta_n^k)$. It follows that

$$\int_{\theta_n^k + 1/2^m}^{\zeta_n^k} \frac{g(u - \theta_n^k)}{1 - G(u - \theta_n^k)}\,du \le A_{m,n}^k[0,\infty)$$
$$\le \int_{\theta_n^k}^{\zeta_n^k + 1/2^m} \frac{g(u - \theta_n^k)}{1 - G(u - \theta_n^k)}\,du$$
$$= -\ln\left(1 - G\left(\zeta_n^k + \frac{1}{2^m} - \theta_n^k\right)\right).$$

The random variable $\zeta_n^k - \theta_n^k$ is distributed according to G because it represents a service time. Hence, $G(\zeta_n^k - \theta_n^k)$ is uniformly distributed in $(0,1)$. Due to the continuity of G, for every ω, there exists a sufficiently large $m_0 = m_0(\omega)$ such that for $m \ge m_0$, $G(\zeta_n^k(\omega) - \theta_n^k(\omega) + 1/2^m) < 1$, so that $-\ln(1 - G(\zeta_n^k(\omega) + \frac{1}{2^m} - \theta_n^k(\omega))) < \infty$. Combining the last four statements, we conclude that for each ω, $A_{m,n}^k(\omega;[0,\infty))$ is finite for all $m \ge m_0$ and, moreover, that as $m \to \infty$, the measure $A_{m,n}^k(\omega;\cdot)$ converges vaguely to the measure that has density

$$\frac{g(u - \theta_n^k(\omega))}{1 - G(u - \theta_n^k(\omega))}\mathbb{1}_{\{\theta_n^k(\omega) < u \le \zeta_n^k(\omega)\}} = h\big(u - \theta_n^k(\omega)\big)\mathbb{1}_{\{\theta_n^k(\omega) < u \le \zeta_n^k(\omega)\}},$$

which is precisely the measure $A_n^k(\omega;\cdot)$. The latter measure does not charge points, and in particular does not charge $u = T(\omega)$. So we conclude that for every ω,

$$\lim_{m\to\infty}\int_0^\infty \mathbb{1}_{[0,T(\omega))}(s)A_{m,n}^k(\omega;ds) = \int_0^\infty \mathbb{1}_{[0,T(\omega))}(s)A_n^k(\omega;ds)$$
$$\leq \int_0^{B-\theta_n^k(\omega)} h(u)\,du,$$

where B is an upper bound on the stopping time T (note that if $M<\infty$ we may restrict our attention to bounded stopping times whose bound B satisfies $B<\theta_n^k+M$). Therefore, the last term is finite due to the local integrability of the hazard rate function on $[0,M)$. The limit above, along with the bounded convergence theorem, then implies the desired convergence

$$\lim_{m\to\infty}\mathbb{E}\left[\sum_{j=0}^\infty \mathbb{1}_{\{\theta_n^k\le j/2^m<T,\,\xi_n^k>j/2^m\}}\frac{G((j+1)/2^m-\theta_n^k)-G(j/2^m-\theta_n^k)}{1-G(j/2^m-\theta_n^k)}\right]$$
$$=\mathbb{E}\left[\lim_{m\to\infty}\int_0^\infty \mathbb{1}_{[0,T)}(u)\,dA_{m,n}^k(u)\right]$$
$$=\mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T)}(u)\,dA_n^k(u)\right]$$
$$=\mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,T]}(u)\,dA_n^k(u)\right],$$

where the last equality uses the continuity of A_n^k. This establishes (5.26). In particular, this shows that for every $t\in[0,\infty)$, $\mathbb{E}[A_{\mathbf{1}}^{(N)}(t)]=\mathbb{E}[D^{(N)}(t)]\le\mathbb{E}[E^{(N)}(t)+X^{(N)}(0)]$. Thus, the lemma follows from Remark 3.1. □

Because $D^{(N)}=Q_{\mathbf{1}}^{(N)}$ and the ages of customers are continuous and, hence, predictable processes, the following (seemingly stronger) result can be immediately deduced from the proof of the last lemma.

COROLLARY 5.5. *For every $N\in\mathbb{N}$ and $\varphi\in\mathcal{C}_b([0,M)\times\mathbb{R}_+)$, the process $A_\varphi^{(N)}$ is the $\{\mathcal{F}_t^{(N)}\}$-compensator of the process $Q_\varphi^{(N)}$. In particular, the process $M_\varphi^{(N)}$ defined by*

$$M_\varphi^{(N)} \doteq Q_\varphi^{(N)} - A_\varphi^{(N)} \tag{5.27}$$

is a local $\{\mathcal{F}_t^{(N)}\}$-martingale.

As usual, let $\overline{Q}_\varphi^{(N)}$, $\overline{A}_\varphi^{(N)}$ and $\overline{M}_\varphi^{(N)}$, respectively, denote the scaled processes $Q_\varphi^{(N)}/N$, $A_\varphi^{(N)}/N$ and $M_\varphi^{(N)}/N$. The following lemma will be used in Section 5.3 to establish tightness of these processes.

LEMMA 5.6. *For every $T < \infty$ and $\varphi \in \mathcal{C}_b([0, M) \times \mathbb{R}_+)$,*

$$\limsup_N \mathbb{E}\big[|\overline{Q}_\varphi^{(N)}(T)|\big] < \infty$$

and $\limsup_N \mathbb{E}[\overline{A}_\varphi^{(N)}(T)] < \infty$. *Also, for $t \in [0, \infty)$ and $N \in \mathbb{N}$,*

$$\lim_{\delta \to 0} \mathbb{E}\big[\overline{D}^{(N)}(t+\delta) - \overline{D}^{(N)}(t)\big] = 0. \tag{5.28}$$

Moreover, for every $\delta > 0$ and interval $\mathcal{Z} = [m + \delta, M)$ with $m \in [0, M - \delta)$,

$$\mathbb{E}\big[\overline{Q}_{\mathbb{1}_\mathcal{Z}}^{(N)}(t+\delta) - \overline{Q}_{\mathbb{1}_\mathcal{Z}}^{(N)}(t) | \mathcal{F}_t^{(N)}\big] \le U(\delta)\overline{v}_t^{(N)}[m, M), \tag{5.29}$$

where $U(\cdot)$ is the renewal function associated with the service distribution G.

PROOF. For every $T < \infty$ and $\varphi \in \mathcal{C}_b([0, M) \times \mathbb{R}_+)$,

$$\mathbb{E}\big[|\overline{A}_\varphi^{(N)}(T)|\big] \le \mathbb{E}\big[\overline{A}_{|\varphi|}^{(N)}(T)\big] = \mathbb{E}\big[\overline{Q}_{|\varphi|}^{(N)}(T)\big],$$

where the last equality is justified by Corollary 5.5. Therefore, the first assertion of the lemma follows from (5.3). For notational conciseness, throughout the rest of this proof we will use $f(t, t+\delta]$ to denote $f(t+\delta) - f(t)$ for any function f, $t \in [0, \infty)$ and $\delta > 0$. Since $\mathbb{1}_\mathcal{Z}$ is only a function of x, we can write

$$Q_{\mathbb{1}_\mathcal{Z}}^{(N)}(t, t+\delta] = \sum_{j=-\langle 1, v_0^{(N)}\rangle + 1}^{K^{(N)}(t+\delta)} \sum_{s \in (t, t+\delta]} \mathbb{1}_{\{d/dt a_j^{(N)}(s-) > 0, a_j^{(N)}(s) = v_j\}} \mathbb{1}_\mathcal{Z}\big(a_j^{(N)}(s)\big),$$

which is simply the number of departures from the N-server system during the time interval $(t, t+\delta]$ by customers whose ages at the time of departure (which equals their service times) lie in the set $\mathcal{Z}$.

We shall bound the departures during the time interval $(t, t+\delta]$ in the N-server system by the departures in another system that is easier to analyze. Consider a modified system in which at time t, there are an infinite number of arrivals (or, equivalently, customers in queue) so that after t, at each station, every time a customer finishes service, a new customer joins. Let $\tilde{D}_1(\delta|x)$ denote the number of departures from a single station in this modified system during the period $(t, t+\delta]$, given that at time t there exists a customer with age x in that station (note that, as the notation reflects, this quantity is independent of t and the choice of station). In fact, the quantity $\tilde{D}_1(\delta|x)$ is simply the number of renewals in the interval $[0, \delta]$ of a delayed renewal process with initial distribution that has density $g_0(y) = g(y+x)/(1 - G(x))$, and inter-renewal distribution G. Thus, as is well known (see, e.g., Theorem 2.4(iii) of [1]), $\mathbb{E}[\tilde{D}_1(\delta|x)]$ is bounded above by $U(\delta)$, where $U(\cdot)$ is the renewal function of a pure (zero-delayed) renewal process that has inter-renewal distribution G (and a renewal at 0).

Let $\tilde{D}(t, t+s]$ be the departure process from the modified system during the interval $(t, t+s]$. At time t, each customer present in the original system is also

present in the modified system and has the same age in both systems. On the other hand, if there are idle servers in the original system at time t, that is, if $N - \langle \mathbf{1}, \nu_t^{(N)} \rangle > 0$, then the modified system has $N - \langle \mathbf{1}, \nu_t^{(N)} \rangle$ servers that have customers with age zero at time t. Thus, $\mu_t^{(N)} \doteq \nu_t^{(N)} + (N - \langle \mathbf{1}, \nu_t^{(N)} \rangle)\delta_0$ represents the age distribution of customers in the modified system at time t. As a result, $\mu_t^{(N)}[0, M) = N$ and a simple monotonicity argument shows that

$$\begin{aligned} \mathbb{E}\big[\overline{D}^{(N)}(t+\delta) - \overline{D}^{(N)}(t) | \mathcal{F}_t^{(N)}\big] &\le \frac{1}{N}\mathbb{E}\big[\tilde{D}(t, t+\delta] | \mathcal{F}_t^{(N)}\big] \\ &= \frac{1}{N}\int_{[0,M)} \mathbb{E}[\tilde{D}_1(\delta|x)]\mu_t^{(N)}(dx) \\ &\le U(\delta). \end{aligned} \tag{5.30}$$

Now, $U(\delta)$ is finite for any finite δ and nondecreasing (see, e.g., Theorem 2.4(i) of [1]). Because $\mathbb{E}[\tilde{D}_1(\delta|x)]$ converges monotonically down to zero as $\delta \to 0$, the bounded convergence theorem shows that for every $N \in \mathbb{N}$,

$$\lim_{\delta \to 0} \int_{[0,M)} \mathbb{E}[\tilde{D}_1(\delta|x)] \frac{1}{N} \mu_t^{(N)}(dx) = 0.$$

Taking expectations of both sides of (5.30) and then sending $\delta \to 0$, the last display and another application of the bounded convergence theorem shows that (5.28) is satisfied for every $N \in \mathbb{N}$.

To establish (5.29), fix $\delta > 0$ and $m \in (0, M - \delta)$. Then any customer whose service time is greater than or equal to $m + \delta$ and who departed the system during the time interval $(t, t + \delta]$ must have been in the system at time t with age greater than or equal to $m > 0$. Thus, the total number of such departures is bounded above by the number of departures in the modified system from stations that had a customer present at time t with age greater than or equal to m. By the same reasoning provided above, this implies that

$$\begin{aligned} \mathbb{E}\big[\overline{Q}^{(N)}_{\mathbb{1}_{[m+\delta,M)}}(t, t+\delta) | \mathcal{F}_t^{(N)}\big] &\le \int_{[m,M)} \mathbb{E}[\tilde{D}_1(\delta|x)]\overline{\nu}_t^{(N)}(dx) \\ &\le U(\delta)\overline{\nu}_t^{(N)}[m, M), \end{aligned}$$

which completes the proof of the lemma. □

We now derive another estimate, which can be viewed as a "pre-limit" analogue of the estimate (4.34) that was obtained for solutions of the age equation in Proposition 4.15.

PROPOSITION 5.7. *Given $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$ and $\varphi \in \mathcal{C}_b([0, M) \times \mathbb{R}_+)$, $\int_0^{\cdot} \langle \ell(\cdot) \times \varphi(\cdot, s), \overline{\nu}_s^{(N)} \rangle \, ds$ is well defined for every $N \in \mathbb{N}$. Moreover, if $\ell \in \mathcal{L}^1[0, M)$, then*

for every $0 \le r \le t < \infty$,

$$\begin{aligned}
&\left| \int_r^t \langle \varphi(\cdot, s)\ell(\cdot), \overline{\nu}_s^{(N)} \rangle \, ds \right| \\
&\qquad \le \|\varphi\|_\infty \big(\overline{X}^{(N)}(0) + \overline{E}^{(N)}(t) \big) \sup_{u \in [0,M)} \int_u^{(u+t-r)\wedge M} |\ell(x)| \, dx.
\end{aligned} \tag{5.31}$$

PROOF. The fact that $\int_0^t \langle \ell(\cdot)\varphi(\cdot, s), \nu_s^{(N)} \rangle \, ds$ equals the right-hand side of (5.24), but with h replaced by ℓ, shows that it is well defined. To obtain the estimate (5.31), manipulating (5.24) with h replaced by ℓ, we obtain for every $N \in \mathbb{N}$ and $0 < r < t < \infty$,

$$\begin{aligned}
&\left| \int_r^t \langle \varphi(\cdot, s)\ell(\cdot), \overline{\nu}_s^{(N)} \rangle \, ds \right| \\
&\qquad \le \frac{\|\varphi\|_\infty}{N} \sum_{j=-\langle 1, \nu_0^{(N)} \rangle + 1}^{0} \int_r^t |\ell(a_j^{(N)}(0) + s)| \mathbb{1}_{\{a_j^{(N)}(0) + s < v_j\}} \, ds \\
&\qquad\quad + \frac{\|\varphi\|_\infty}{N} \sum_{j=1}^{K^{(N)}(r)} \int_r^t |\ell(s - \alpha_j^{(N)})| \mathbb{1}_{\{s - \alpha_j^{(N)} \le v_j\}} \, ds \\
&\qquad\quad + \frac{\|\varphi\|_\infty}{N} \sum_{j=K^{(N)}(r)+1}^{K^{(N)}(t)} \int_{\alpha_j^{(N)}}^t |\ell(s - \alpha_j^{(N)})| \mathbb{1}_{\{s - \alpha_j^{(N)} \le v_j\}} \\
&\qquad \le \|\varphi\|_\infty \big(\langle \mathbf{1}, \overline{\nu}_0^{(N)} \rangle + \overline{K}^{(N)}(t) \big) \sup_{u \in [0,M)} \int_u^{(u+t-r)\wedge M} |\ell(x)| \, dx,
\end{aligned} \tag{5.32}$$

where the last inequality uses the fact that almost surely $v_j < M$ and $\alpha_j^{(N)} \in [r, t]$ for $j = K^{(N)}(r) + 1, \ldots, K^{(N)}(t)$. The estimate (5.31) then follows from the above display, (2.5), (2.6), the nonnegativity of $\overline{X}^{(N)}$ and the fact that $\overline{\nu}_s^{(N)}$ is a sub-probability measure for every s. □

In the next lemma, these estimates are used to obtain some convergence results, which will in turn be used to prove tightness of the sequence of state processes in Section 5.3. Assumptions (5.33) and (5.34) imposed in the lemma below are shown to follow from Assumption 1(3) in Lemma 5.12.

LEMMA 5.8. *Suppose that the limit*

$$\lim_{m \uparrow M} \sup \mathbb{E}\big[\overline{\nu}_0^{(N)}(m, M) \big] = 0 \tag{5.33}$$

holds, and if $M < \infty$, *then*

$$\lim_{m\uparrow M} \sup_N \mathbb{E}\left[\int_{[0,m)} \frac{1-G(m)}{1-G(x)} \overline{\nu}_0^{(N)}(dx)\right] = 0 \tag{5.34}$$

is also satisfied. Then the following three properties hold:

(1) *For* $t \in [0,\infty)$,

$$\lim_{m\uparrow M} \sup_N \mathbb{E}\left[\int_0^t \left(\int_{[m,M)} h(x)\overline{\nu}_s^{(N)}(dx)\right) ds\right] = 0.$$

(2) *For every* $\varphi \in \mathcal{C}_b([0,M)\times\mathbb{R}_+)$ *and* $T \in [0,\infty)$,

$$\lim_{\delta\to 0} \limsup_N \mathbb{E}\Big[\sup_{t\in[0,T]} \big(\overline{A}_\varphi^{(N)}(t+\delta) - \overline{A}_\varphi^{(N)}(t)\big)\Big] = 0$$

and for $t \in [0,\infty)$,

$$\lim_{\delta\to 0} \limsup_N \mathbb{E}\big[\overline{Q}_\varphi^{(N)}(t+\delta) - \overline{Q}_\varphi^{(N)}(t)\big] = 0.$$

(3) *Given* $m < M$ *and any sequence of subsets* $\mathcal{H}_n \subset [0,m]$ *such that the Lebesgue measure of the set* $\mathcal{H}_n$ *goes to zero as* $n\to\infty$, *we have for every* $T \in [0,\infty)$,

$$\lim_{n\to\infty} \limsup_N \mathbb{E}\Big[\sup_{t\in[0,T]} \overline{A}_{\mathbb{1}_{\mathcal{H}_n}}^{(N)}(t)\Big] = 0. \tag{5.35}$$

PROOF. As in the last two proofs, in this proof too we will use $f(t,t+\delta]$ to denote $f(t+\delta) - f(t)$ for any function f, $t \in [0,\infty)$ and $\delta > 0$. We shall divide the proof of the first property into two cases.

Case 1. $M = \infty$. We start by proving a preliminary result, (5.36) below. For $m, \Delta, s \in [0,\infty)$, let $f_{m,\Delta,s} \in \mathcal{C}_b(\mathbb{R}_+)$ be such that

$$\mathbb{1}_{[2m+\Delta+s,\infty)} \le f_{m,\Delta,s} \le \mathbb{1}_{[m+\Delta+s,\infty)}.$$

By Corollary 5.5 and (5.29), we have for every $N \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\big[\overline{A}_{f_{m,\Delta,s}}^{(N)}(s,s+\Delta]|\mathcal{F}_s^{(N)}\big] &= \mathbb{E}\big[\overline{Q}_{f_{m,\Delta,s}}^{(N)}(s,s+\Delta]|\mathcal{F}_s^{(N)}\big] \\ &\le \mathbb{E}\big[\overline{Q}_{\mathbb{1}_{[m+\Delta+s,\infty)}}^{(N)}(s,s+\Delta]|\mathcal{F}_s^{(N)}\big] \\ &\le U(\Delta)\overline{\nu}_s^{(N)}[m+s,\infty).\end{aligned}$$

Taking expectations of both sides, we see that

$$\mathbb{E}\big[\overline{A}_{f_{m,\Delta,s}}^{(N)}(s,s+\Delta]\big] \le U(\Delta)\mathbb{E}\big[\overline{\nu}_s^{(N)}[m+s,\infty)\big]. \tag{5.36}$$

We now show how the first property (in the case $M = \infty$) follows from the above estimate. Fix $t \in [0, \infty)$, choose $m > t$ and let $\tilde{m} \doteq (m - t)/2$. Then we have

$$\begin{aligned}\mathbb{E}\Big[\int_0^t \Big(\int_{[m,\infty)} h(x)\overline{\nu}_s^{(N)}(dx)\Big) ds\Big] &= \mathbb{E}\Big[\int_0^t \Big(\int_{[2\tilde{m}+t,\infty)} h(x)\overline{\nu}_s^{(N)}(dx)\Big) ds\Big] \\ &\le \mathbb{E}\big[\overline{A}^{(N)}_{f_{\tilde{m},t,0}}(t)\big] \\ &\le U(t)\mathbb{E}\big[\overline{\nu}_0^{(N)}[\tilde{m}, \infty)\big],\end{aligned}$$

where the last inequality is justified by the estimate (5.36), with m, Δ and s replaced by $\tilde{m}, t$ and 0, respectively. Taking the supremum of both sides over N, and then sending $m \to \infty$ (in which case $\tilde{m} \to \infty$), relation (5.33) ensures that property (1) holds for the case $M = \infty$.

Case 2. $M < \infty$. In this case, for $m < M$, by Corollary 5.5 and the fact that $\mathbb{1}_{(m,M)}(a_j^{(N)}(\cdot))$ is left-continuous and hence predictable, we have

$$\begin{aligned}&\mathbb{E}\Big[\int_0^t \Big(\int_{(m,M)} h(x)\overline{\nu}_s^{(N)}(dx)\Big) ds\Big] \\ &\quad= \mathbb{E}\big[\overline{A}^{(N)}_{\mathbb{1}_{(m,M)}}(t)\big] = \mathbb{E}\big[\overline{Q}^{(N)}_{\mathbb{1}_{(m,M)}}(t)\big] \\ &\quad= \frac{1}{N}\mathbb{E}\Bigg[\sum_{s\in[0,t]} \sum_{j=-\langle 1,\nu_0^{(N)}\rangle+1}^{K^{(N)}(t)} \mathbb{1}_{\{d/dt a_j^{(N)}(s-)>0, a_j^{(N)}(s)=v_j\}} \mathbb{1}_{(m,M)}\big(a_j^{(N)}(s)\big)\Bigg] \\ &\quad\le \mathbb{E}\big[\overline{\nu}_0^{(N)}(m, M)\big] + \frac{1}{N}\mathbb{E}\Bigg[\sum_{j=-\langle 1,\nu_0^{(N)}\rangle+1}^{0} \mathbb{1}_{\{a_j^{(N)}(0)\le m\}} \mathbb{1}_{\{v_j\in(m,M)\}}\Bigg] \\ &\qquad+ \frac{1}{N}\mathbb{E}\Bigg[\sum_{j=1}^{E^{(N)}(t)} \mathbb{1}_{\{v_j\in(m,M)\}}\Bigg].\end{aligned}$$

Conditioning on $\mathcal{F}_0^{(N)}$ and using the fact that $a_j^{(N)}(0)$, $j = -\langle \nu_0^{(N)}, \mathbf{1}\rangle, \ldots, 0$ and, hence, $\nu_0^{(N)}$ are measurable with respect to $\mathcal{F}_0^{(N)}$, we see that the second term on the right-hand side can be rewritten as

$$\begin{aligned}&\frac{1}{N}\mathbb{E}\Bigg[\sum_{j=-\langle 1,\nu_0^{(N)}\rangle+1}^{0} \mathbb{1}_{\{a_j^{(N)}(0)\le m\}} \mathbb{1}_{\{v_j\in(m,M)\}}\Bigg] \\ &\qquad= \frac{1}{N}\mathbb{E}\Bigg[\sum_{j=-\langle 1,\nu_0^{(N)}\rangle+1}^{0} \mathbb{1}_{\{a_j^{(N)}(0)\le m\}} \mathbb{E}\big[\mathbb{1}_{\{v_j\in(m,M)\}} | a_j^{(N)}(0)\big]\Bigg]\end{aligned}$$

$$= \frac{1}{N}\mathbb{E}\Bigg[\sum_{j=-\langle 1,\nu_0^{(N)}\rangle+1}^{0} \mathbb{1}_{\{a_j^{(N)}(0)\le m\}}\frac{1-G(m)}{1-G(a_j^{(N)}(0))}\Bigg]$$

$$= \mathbb{E}\Bigg[\int_{[0,m]}\frac{1-G(m)}{1-G(x)}\overline{\nu}_0^{(N)}(dx)\Bigg].$$

On the other hand, the independence of the arrival process and the service requirements of the customers implies that the third term can be simplified to

$$\frac{1}{N}\mathbb{E}\Bigg[\sum_{j=1}^{E^{(N)}(t)}\mathbb{1}_{\{v_j\in[m,M)\}}\Bigg] = \big(1-G(m)\big)\mathbb{E}\big[\overline{E}^{(N)}(t)\big].$$

Combining the last three displays, we conclude that

$$\begin{aligned}\mathbb{E}\Bigg[\int_0^t\Bigg(\int_{(m,M)}h(x)\overline{\nu}_s^{(N)}(dx)\Bigg)ds\Bigg]&\\ \le \mathbb{E}\big[\overline{\nu}_0^{(N)}(m,M)\big] + \mathbb{E}\Bigg[\int_{[0,m]}\frac{1-G(m)}{1-G(x)}\overline{\nu}_0^{(N)}(dx)\Bigg]&\\ + \big(1-G(m)\big)\mathbb{E}\big[\overline{E}^{(N)}(t)\big].&\end{aligned}$$

Taking the supremum of both sides over N and then sending $m\to M$, (5.33), (5.34) and Assumption 1(1) ensure that property 1 is satisfied.

We now turn to the proof of property 2. Fix $\varphi\in\mathcal{C}_b([0,M)\times\mathbb{R}_+)$ and $T\in[0,\infty)$. For any $m<M$ and $t\in[0,T]$, we have

$$\begin{aligned}\mathbb{E}\Big[\sup_{t\in[0,T]}\overline{A}_{\varphi}^{(N)}(t,t+\delta]\Big] &\le \mathbb{E}\Big[\sup_{t\in[0,T]}\overline{A}_{\varphi\mathbb{1}_{[0,m]}}^{(N)}(t,t+\delta]\Big]\\ &\quad+\mathbb{E}\Big[\sup_{t\in[0,T]}\overline{A}_{\varphi\mathbb{1}_{(m,M)}}^{(N)}(t,t+\delta]\Big].\end{aligned}\tag{5.37}$$

However, applying (5.31) with $\ell = h\mathbb{1}_{[0,m]}$, and r and t replaced by t and $t+\delta$, respectively, then taking the supremum over $t\in[0,T]$, next the expectation and finally the limit superior over N, we obtain

$$\limsup_{N\to\infty}\mathbb{E}\Big[\sup_{t\in[0,T]}\overline{A}_{\varphi\mathbb{1}_{[0,m]}}^{(N)}(t,t+\delta]\Big] \le \|\varphi\|_\infty C(T+\delta)\sup_{u\in[0,m]}\int_u^{(u+\delta)\wedge m}h(x)\,dx,$$

where $C(T+\delta)=\mathbb{E}[\overline{X}(0)+\overline{E}(T+\delta)]<\infty$ by properties 1 and 2 of Assumption 1. The right-hand side goes to zero as $\delta\to 0$ because h is locally integrable. As a result, we have

$$\lim_{\delta\to 0}\limsup_{N\to\infty}\mathbb{E}\Big[\sup_{t\in[0,T]}\overline{A}_{\varphi\mathbb{1}_{[0,m]}}^{(N)}(t,t+\delta]\Big]=0.$$

Taking first the limit superior of (5.37), as $N\to\infty$, next sending $\delta\to 0$ and then taking the limit as $m\uparrow M$, the first term on the right-hand side vanishes due to the

last display, whereas the second term goes to zero by property 1. This proves the first relation of property 2. The second relation of property 2 follows trivially from the first on account of Corollary 5.5.

Once again considering (5.31), this time with $\varphi = \mathbf{1}$, $\ell_n = h\mathbb{1}_{\mathcal{H}_n} = h\mathbb{1}_{\mathcal{H}_n \cap [0,m]}$ and $r = 0$, taking first the supremum over $t \in [0, T]$, next expectations and then the limit superior, as $N \to \infty$, we obtain

$$\limsup_{N\to\infty} \mathbb{E}\Big[\sup_{t\in[0,T]} \overline{A}^{(N)}_{\mathbb{1}_{\mathcal{H}_n}}(t)\Big] \le C(T) \int_{\mathcal{H}_n \cap [0,m]} h(x)\,dx,$$

with $C(T) = \limsup_N \mathbb{E}[\overline{X}^{(N)}(0) + \overline{E}(T)] < \infty$, where the finiteness is a consequence of properties 1 and 2 of Assumption 1. Sending $n \to \infty$, the local integrability of h and the fact that the Lebesgue measure of $\mathcal{H}_n$ converges to zero as $n \to \infty$ show that the right-hand side above tends to zero. This proves the last property of the lemma. □

The local martingale $\overline{M}^{(N)}_\varphi$ has a well-defined predictable quadratic variation process $\langle \overline{M}^{(N)}_\varphi \rangle$ because $\overline{M}^{(N)}_\varphi(0) = 0$ and $M^{(N)}_\varphi$ has bounded jumps (see, e.g., statement (4.1) of Section I of [10]). We now show that the sequence of predictable quadratic variation processes $\langle \overline{M}^{(N)}_\varphi \rangle$, $N \in \mathbb{N}$, converges to zero as $N \to \infty$.

LEMMA 5.9. *For every $\varphi \in \mathcal{C}_b([0, M) \times \mathbb{R}_+)$ and $t \in [0, \infty)$,*

$$\lim_{N\to\infty} \mathbb{E}[\langle \overline{M}^{(N)}_\varphi \rangle(t)] = 0. \tag{5.38}$$

Consequently, $\overline{M}^{(N)}_\varphi \Rightarrow \mathbf{0}$ as $N \to \infty$.

PROOF. Because $M^{(N)}_\varphi$ is a compensated sum of jumps with a continuous compensator $A^{(N)}_\varphi$, $M^{(N)}_\varphi$ does not have any predictable jump times, that is, $\Delta M^{(N)}_\varphi(T) = 0$ for every predictable time T (see, e.g., Corollary 1.19 of Section II and Definition 2.25 of Section I in [10]). Therefore, by Proposition 2.29 of Section II in [10], the predictable quadratic variation of the martingale is given by

$$\langle M^{(N)}_\varphi \rangle(t) = \int_0^t \left(\int_{[0,M)} \varphi^2(x,s) h(x) \nu^{(N)}_s(dx) \right) ds.$$

This means that the scaled process $\overline{M}^{(N)}_\varphi$ has predictable quadratic variation

$$\langle \overline{M}^{(N)}_\varphi \rangle(t) = \frac{1}{N^2} \langle M^{(N)}_\varphi \rangle(t) = \frac{1}{N} \left[\int_0^t \left(\int_{[0,M)} \varphi^2(x,s) h(x) \overline{\nu}^{(N)}_s(dx) \right) ds \right],$$

which implies that for any $t \in [0, \infty)$,

$$\langle \overline{M}^{(N)}_\varphi \rangle(t) \le \frac{\|\varphi\|^2_\infty}{N} \overline{A}^{(N)}_{\mathbf{1}}(t).$$

Noting that $\mathbb{E}[\overline{A}_{\mathbf{1}}^{(N)}(t)] = \mathbb{E}[\overline{D}^{(N)}(t)]$ by Lemma 5.4, the first assertion of Lemma 5.6 shows that $\sup_{N\in\mathbb{N}} \mathbb{E}[\overline{A}_{\mathbf{1}}^{(N)}(t)] < \infty$. Thus, taking first expectations and then limits as $N \to \infty$ in the last display, we obtain the first assertion of the lemma. To show that $\overline{M}_{\varphi}^{(N)} \Rightarrow \mathbf{0}$ as $N \to \infty$, we note that by Doob's lemma, for any $\lambda > 0$

$$\mathbb{P}\Big(\sup_{s\in[0,T]} \big|\overline{M}_{\varphi}^{(N)}(s)\big| > \lambda \Big) \le \frac{\mathbb{E}[\langle \overline{M}_{\varphi}^{(N)} \rangle(T)]}{\lambda^2},$$

which converges to 0 as $N \to \infty$ by the first assertion. Because this is true for all $\lambda > 0$, this completes the proof of the lemma. □

5.3. *Proof of relative compactness.* We now establish the relative compactness of the sequence of scaled state processes $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)})\}$, as well as of several of the auxiliary processes. For this, it will be convenient to use Kurtz's criteria for relative compactness of processes $\{Y^{(N)}\}$ with sample paths in $\mathcal{D}_{\mathbb{R}}[0,\infty)$.

Kurtz's criteria.

K1. For every rational $t \ge 0$,

$$\lim_{R\to\infty} \sup_N \mathbb{P}\big(\big|Y^{(N)}(t)\big| > R\big) = 0; \tag{5.39}$$

K2. For each $t > 0$, there exists $\beta > 0$ such that

$$\lim_{\delta\to 0} \sup_N \mathbb{E}\big[\big|Y^{(N)}(t+\delta) - Y^{(N)}(t)\big|^{\beta}\big] = 0. \tag{5.40}$$

The sufficiency of these conditions for relative compactness follows from Theorem 3.8.6 of [5] (condition K1 corresponds to condition (a) of Theorem 3.7.2 in [5], and condition K2 follows from condition (b) of Theorem 3.8.6 and Remark 3.8.7 in [5]).

LEMMA 5.10. *Suppose Assumption 1 holds. Then the sequences* $\{\overline{Q}_{\varphi}^{(N)}\}$, $\{\overline{A}_{\varphi}^{(N)}\}$ *and* $\{\overline{M}_{\varphi}^{(N)}\}$, $\varphi \in \mathcal{C}_b([0,M)\times\mathbb{R}_+)$, *the sequences* $\{\overline{X}^{(N)}\}$ *and* $\{\langle \mathbf{1}, \overline{\nu}^{(N)}\rangle\}$, *and the sequences* $\{\langle f, \overline{\nu}_{\cdot}^{(N)}\rangle\}$, $f \in \mathcal{C}_b^1[0,M)$, *are all relatively compact in* $\mathcal{D}_{\mathbb{R}}[0,\infty)$.

PROOF. Since we are working on Polish spaces, by Prohorov's theorem the notions of relative compactness and tightness are equivalent. Fix $T < \infty$ and $\varphi \in \mathcal{C}_b([0,M)\times\mathbb{R}_+)$. The fact that $\{\overline{A}_{\varphi}^{(N)}\}$ and $\{\overline{Q}_{\varphi}^{(N)}\}$ satisfy condition K1 is easily deduced from the bounds $\sup_N \mathbb{E}[|\overline{Q}_{\varphi}^{(N)}(T)|] < \infty$ and $\sup_N \mathbb{E}[|\overline{A}_{\varphi}^{(N)}(T)|] < \infty$ that were proved in Lemma 5.6. In addition, from Lemma 5.8(2) it follows that for every $\varphi \in \mathcal{C}_b([0,M)\times\mathbb{R})$, the sequences $\{\overline{A}_{\varphi}^{(N)}\}$ and $\{\overline{Q}_{\varphi}^{(N)}\}$ satisfy criterion K2

(with $\beta = 1$), and thus are relatively compact. Finally, K1 and K2 for $\{\overline{M}_\varphi^{(N)}\}$ follow from K1 and K2 for $\{\overline{A}_\varphi^{(N)}\}$ and $\{\overline{Q}_\varphi^{(N)}\}$.

By Assumption 1 and the result just proved above, the sequences $\{\overline{E}^{(N)}\}$ and $\{\overline{X}^{(N)}(0)\}$ and $\{\overline{D}^{(N)}\}$ are tight and, therefore, relatively compact. Hence, by Theorem 3.7.2 of [5], they satisfy (5.39). The elementary bound

$$\langle \mathbf{1}, \overline{\nu}_t^{(N)} \rangle \le \overline{X}^{(N)}(t) \le \overline{X}^{(N)}(0) + \overline{E}^{(N)}(t), \qquad t \in [0, \infty),$$

then shows that the sequences $\{\overline{X}^{(N)}\}$ and $\{\langle \mathbf{1}, \overline{\nu}^{(N)} \rangle\}$ also satisfy condition K1. To prove the relative compactness of these sequences, we will use a slightly different set of criteria, namely K1 above and condition (b) of Theorem 3.7.2 of [5], which is expressed in terms of the modulus of continuity $w'(f, \delta, T)$ of a function f (see (3.6.2) of [5] for a precise definition of w'). For every $0 \le s \le t < \infty$, from (5.5) it is clear that

$$\big|\overline{X}^{(N)}(t) - \overline{X}^{(N)}(s)\big| \le \big|\overline{E}^{(N)}(t) - \overline{E}^{(N)}(s)\big| \vee \big|\overline{D}^{(N)}(t) - \overline{D}^{(N)}(s)\big|,$$

and the complementarity condition (5.6) shows that

$$\big|\langle \mathbf{1}, \overline{\nu}_t^{(N)} \rangle - \langle \mathbf{1}, \overline{\nu}_s^{(N)} \rangle\big| \le \big|\big[1 - \overline{X}^{(N)}(t)\big]^+ - \big[1 - \overline{X}^{(N)}(s)\big]^+\big| \le \big|\overline{X}^{(N)}(t) - \overline{X}^{(N)}(s)\big|.$$

From this it is easy to see that for every $N \in \mathbb{N}$, $\delta > 0$ and $T < \infty$,

$$w'\big(\langle \mathbf{1}, \overline{\nu}^{(N)} \rangle, \delta, T\big) \vee w'\big(\overline{X}^{(N)}, \delta, T\big) \le w'\big(\overline{E}^{(N)}, \delta, T\big) \vee w'\big(\overline{D}^{(N)}, \delta, T\big).$$

The relative compactness of $\{\overline{X}^{(N)}\}$ and $\{\langle \mathbf{1}, \overline{\nu}^{(N)} \rangle\}$ is then a direct consequence of the above estimate, the relative compactness of $\{\overline{E}^{(N)}\}$ and $\{\overline{D}^{(N)}\}$ and Theorem 3.7.2 of [5].

Now, let $f \in \mathcal{C}_b^1[0, M)$. We shall prove the relative compactness of the sequence $\{\langle f, \overline{\nu}_\cdot^{(N)} \rangle\}$. First, substituting $\varphi = f$ (as usual, interpreting f as a function on $[0, M) \times \mathbb{R}_+$ that depends only on the first variable and noting that then the continuous differentiability of f in the first variable trivially guarantees that $f \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$) in the N-server equation (5.4), and then dividing the equation by N, we obtain for any $t \in [0, \infty)$,

$$\begin{aligned}
\langle f, \overline{\nu}_t^{(N)} \rangle - \langle f, \overline{\nu}_0^{(N)} \rangle &= \int_0^t \langle f', \overline{\nu}_s^{(N)} \rangle ds - \overline{Q}_f^{(N)}(t) + f(0)\overline{K}^{(N)}(t) \\
&= \int_0^t \langle f', \overline{\nu}_s^{(N)} \rangle ds - \overline{Q}_f^{(N)}(t) \\
&\quad + f(0)\big[\overline{Q}_{\mathbf{1}}^{(N)}(t) + \langle \mathbf{1}, \overline{\nu}_t^{(N)} \rangle - \langle \mathbf{1}, \overline{\nu}_0^{(N)} \rangle\big],
\end{aligned}$$

where the last equality uses the relation (2.6). Thus, to show that $\{\langle f, \overline{\nu}^{(N)} \rangle\}$ is relatively compact, it suffices to show that $\{\langle f, \overline{\nu}_0^{(N)} \rangle\}$ and the sequences associated with each of the three terms on the right-hand side of the last display are relatively compact. The relative compactness of the last two terms is a direct result of the relative compactness of $\{\overline{Q}_f^{(N)}\}$, $\{\overline{Q}_{\mathbf{1}}^{(N)}\}$ and $\{\langle \mathbf{1}, \overline{\nu}^{(N)} \rangle\}$ proved above and the relative

compactness of the sequence $\{\langle \mathbf{1}, \overline{\nu}_0^{(N)}\rangle\}$, which holds due to Assumption 1(3). In addition, since $\overline{\nu}^{(N)}$ is a sub-probability measure for every $N \in \mathbb{N}$, the first term is uniformly bounded by $\|f'\|_\infty t$ and, moreover,

$$\int_t^{t+u} |\langle f', \overline{\nu}_s^{(N)}\rangle|\, ds \le \|f'\|_\infty u.$$

This verifies Kurtz' criteria K1 and K2 (with $\beta = 1$), for the sequence associated with the first term and, hence, establishes its relative compactness. □

We now show that several sequences of measure-valued processes associated with the many-server model are relatively compact. Given a complete separable metric space S, the space $\mathcal{M}_F(S)$, equipped with the topology of weak convergence, is also a complete separable metric space. Therefore, Jakubowski's criteria (see, e.g., Theorem 4.6 of [11]) for tightness, summarized below, can be applied to establish relative compactness on $\mathcal{D}_{\mathcal{M}_F(S)}[0,\infty)$, with $S = [0, M)$ and $S = [0, M) \times \mathbb{R}_+$.

Jakubowski's criteria. A sequence $\{\pi^{(N)}\}$ of $\mathcal{D}_{\mathcal{M}_F(S)}[0,\infty)$-valued random elements defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is tight if and only if the following two conditions are satisfied:

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

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

This is referred to as the *compact containment condition*.

J2. There exists a family $\mathbb{F}$ of real-valued continuous functions H on $\mathcal{M}_F(S)$ that separates points in $\mathcal{M}_F(S)$ and is closed under addition such that $\{\tilde{\nu}^{(N)}\}$ is $\mathbb{F}$-weakly tight, that is, for every $H \in \mathbb{F}$, the sequence $\{H(\pi_s^{(N)}), s \in [0,\infty)\}$, $N \in \mathbb{N}$, is tight in $\mathcal{D}_{\mathbb{R}}[0,\infty)$.

REMARK 5.11. Consider the family of real-valued functions $\mathbb{F}$ on $\mathcal{M}_1(S)$ given by

$$\mathbb{F} \doteq \{H : \exists f \in \mathcal{C}_b^1[0, M) \text{ such that } H(\mu) = \langle f, \mu\rangle \ \forall \mu \in \mathcal{M}_1(S)\}.$$

Every function in $\mathbb{F}$ is clearly continuous with respect to the weak topology on $\mathcal{M}_1(S)$ and the class $\mathbb{F}$ is trivially closed with respect to addition. Moreover, $\mathbb{F}$ clearly separates points in $\mathcal{M}_1(S)$.

We start by establishing the relative compactness of the sequence of measure-valued processes $\{\overline{\nu}^{(N)}\}$.

LEMMA 5.12. *Suppose Assumption 1 holds. Then the sequence* $\{\overline{\nu}^{(N)}\}$ *is relatively compact. Moreover, the limits* (5.33) *and* (5.34) *hold.*

PROOF. By Lemma 5.10 and Remark 5.11, $\{\overline{\nu}^{(N)}\}$ satisfies Jakubowski's criterion J2. Therefore, it suffices to show that $\{\overline{\nu}^{(N)}\}$ satisfies Jakubowski's criterion J1. Define $\tilde{\nu}^{(N)} \doteq (1 - \langle \mathbf{1}, \overline{\nu}^{(N)} \rangle)\delta_0 + \overline{\nu}^{(N)}$, where δ_0 is the Dirac mass at zero. Then $\tilde{\nu}_s^{(N)}$ is a probability measure on $[0, M)$ for every $s \in [0, \infty)$ and, since $1 - \langle \mathbf{1}, \overline{\nu}_s^{(N)} \rangle \in [0, 1]$, to prove the lemma it clearly suffices to show that $\tilde{\nu}^{(N)}$ satisfies Jakubowski's criterion J1. We split the proof of the latter into two cases, depending on the value of M.

Case 1. $M = \infty$. By Assumption 1(3) and the complementarity condition (5.6), there exists a set $\tilde{\Omega}$ of measure 1 such that for all $\omega \in \tilde{\Omega}$, $\overline{\nu}_0^{(N)}(\omega)$ converges weakly, as $N \to \infty$, to a sub-probability measure $\overline{\nu}_0(\omega)$, which in turn implies that $1 - \overline{\nu}_0^{(N)}(\omega)$ converges to $1 - \langle \mathbf{1}, \overline{\nu}_0(\omega) \rangle$. Fix $\omega \in \tilde{\Omega}$. Then by Prohorov's theorem (see Section 3.2 of [5]), the sequence $\{\tilde{\nu}_0^{(N)}(\omega), N \in \mathbb{N}\}$ must be tight. Hence, for every $\varepsilon > 0$, the positive random variable $r(\omega, \varepsilon)$ defined by

$$r(\omega, \varepsilon) \doteq \sup_N \inf\{a : \tilde{\nu}_0^{(N)}(\omega)[a, \infty) < \varepsilon\}$$

is finite. Note that we then have

$$\tilde{\nu}_0^{(N)}(\omega)(r(\omega, \varepsilon), \infty) < \varepsilon \qquad \text{for all } N \in \mathbb{N}.$$

Since $r(\omega, 1/n) < \infty$ for every $\omega \in \tilde{\Omega}$ and $n \in \mathbb{N}$, there exists a sequence $r(n)$ that converges to infinity as $n \to \infty$, and is such that $\mathbb{P}(\omega : r(\omega, 1/n) > r(n)) \le 2^{-n}$. Define $A_n \doteq \{\omega : r(\omega, 1/n) > r(n)\}$. By the Borel–Cantelli lemma, almost surely A_n occurs only finitely often. Furthermore, $\mathbb{P}(\bigcup_{n \ge N_0} A_n) \le 2^{-N_0+1}$ for every $N_0 \in \mathbb{N}$. Now, fix $T < \infty$ and note that because the age process of each customer in service increases linearly, for every $n \in \mathbb{N}$ and $t \in [0, T]$,

$$\left\{\omega : \tilde{\nu}_0^{(N)}(\omega)(r(n), \infty) \le \frac{1}{n}\right\} \subseteq \left\{\omega : \tilde{\nu}_0^{(N)}(\omega)\big(r(n) - t + T, \infty\big) \le \frac{1}{n}\right\}$$
$$\subseteq \left\{\omega : \tilde{\nu}_t^{(N)}(\omega)\big(r(n) + T, \infty\big) \le \frac{1}{n}\right\}.$$

Thus, given $\eta > 0$, now define

$$\mathcal{K}_{\eta,T} \doteq \left\{\mu \in \mathcal{M}_1(\mathbb{R}_+) : \mu\big(r(n) + T, \infty\big) \le \frac{1}{n} \text{ for all } n > N_0(\eta)\right\},$$

where we choose $N_0(\eta) \doteq -\lceil \ln \eta / \ln 2 \rceil$ so that $2^{-N_0+1} < \eta$. Then observe that

$$\inf_{C \subset \mathbb{R}_+ : C \text{ compact}} \sup_{\mu \in \mathcal{K}_{\eta,T}} \mu(C^c) \le \inf_{n > N_0(\eta)} \sup_{\mu \in \mathcal{K}_{\eta,T}} \mu\big(r(n) + T, \infty\big) = 0.$$

Therefore, another application of Prohorov's theorem shows that $\mathcal{K}_{\eta,T}$ is a relatively compact subset of $\mathcal{M}_1(\mathbb{R}_+)$ (equipped with the Prohorov metric). Let $\overline{\mathcal{K}}_{\eta,T}$

be its closure in the Prohorov metric. Then for every $N \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}\big(\tilde{\nu}_t^{(N)} \in \overline{\mathcal{K}}_{\eta,T} \text{ for every } t \in [0,T]\big) &\geq \mathbb{P}\left(\tilde{\nu}_0^{(N)}(r(n),\infty) \leq \frac{1}{n} \text{ for every } n > N_0(\eta)\right) \geq 1 - 2^{-N_0+1} \\ &\geq 1 - \eta,\end{aligned}$$

which proves the compact containment condition when $M = \infty$.

In addition, this also shows that (5.33) holds. Indeed, if $\eta > 0$ and $N_1(\eta) \in \mathbb{N}$ satisfies $N_1(\eta) \geq N_0(\eta) \vee [1/\eta]$, then the last display implies that for every $m \geq r(N_1(\eta))$,

$$\inf \mathbb{P}\left(\tilde{\nu}_0^{(N)}[m,\infty) \leq \frac{1}{N_1(\eta)}\right) \geq 1 - \eta,$$

which in turn shows that for every for every $m \geq r(N_1(\eta))$,

$$\sup \mathbb{E}\big[\overline{\nu}_0^{(N)}[m,\infty)\big] \leq \frac{1}{N_1(\eta)} + \eta.$$

The result then follows by sending first $m \to \infty$ and then $\eta \to 0$.

Case 2. $M < \infty$. We start by establishing (5.33) and (5.34) using an argument similar to that used in Case 1 to prove (5.33). The almost sure weak convergence of $\overline{\nu}_0^{(N)}$ to $\overline{\nu}_0$ in $\mathcal{M}_{\leq 1}[0,M)$ implies that the sequence $\{r(n)\}$ considered in Case 1 can be taken strictly smaller than M and converging to M. Defining $N_1(\delta)$ as in Case 1, we see that for $m \geq r(N_1(\delta))$,

$$\sup \mathbb{E}\big[\overline{\nu}_0^{(N)}[m,M)\big] \leq \frac{1}{N_1(\eta)} + \eta,$$

and the result follows as above by sending first $m \to M$ and then $\eta \to 0$. The limit (5.34) follows from the weak convergence of $\overline{\nu}_0^{(N)}$ to $\overline{\nu}_0$ and the fact that $(1 - G(L))/(1 - G(x))$ is bounded (by 1) and continuous on $[0, L)$.

It only remains to show that the compact containment condition is satisfied when $M < \infty$. For this, we need to show that for every $\varepsilon > 0$, $\eta > 0$ we can find $m(\varepsilon) < M$ so that

$$\inf_N \mathbb{P}\big(\tilde{\nu}_t^{(N)}[0, m(\varepsilon)] > 1 - \varepsilon \text{ for every } t \in [0,T]\big) > 1 - \eta.$$

However, for any $m < M$, we have

$$\begin{aligned}\mathbb{P}\big(\tilde{\nu}_t^{(N)}(m,M) > \varepsilon \text{ for some } t \in [0,T]\big) &\leq \mathbb{P}\big(\overline{Q}^{(N)}_{\mathbb{1}_{(m,M)}}(T+M) > \varepsilon\big) \\ &\leq \frac{\mathbb{E}[\overline{Q}^{(N)}_{\mathbb{1}_{(m,M)}}(T+M)]}{\varepsilon} \\ &= \frac{\mathbb{E}[\overline{A}^{(N)}_{\mathbb{1}_{(m,M)}}(T+M)]}{\varepsilon},\end{aligned}$$

where the last equality follows from Corollary 5.5. Using now Lemma 5.8(1) [which is justified since we have already established (5.33) and (5.34)], one can find $m(\varepsilon)$ close enough to M to make the supremum over N of the right-hand side above smaller than η, thus yielding the desired result. $\square$

For all $N \in \mathbb{N}$ and $t \in [0, \infty)$, from (5.2), (5.25) and the fact that $\overline{Q}_{\mathbf{1}}^{(N)}(t)$ and $\overline{A}_{\mathbf{1}}^{(N)}(t)$ are a.s. finite, it immediately follows that a.s., the linear functionals $\overline{\mathcal{Q}}^{(N)}(t) : \varphi \mapsto \overline{Q}_{\varphi}^{(N)}(t)$ and $\overline{\mathcal{A}}^{(N)}(t) : \varphi \mapsto \overline{A}_{\varphi}^{(N)}(t)$ on $\mathcal{C}_c([0, M) \times \mathbb{R}_+)$ are finite nonnegative Radon measures on $[0, M) \times \mathbb{R}_+$ (see Section 1.2.2 for a characterization of Radon measures as linear functionals). In other words, for every $\varphi \in \mathcal{C}_c([0, M) \times \mathbb{R}_+)$ the integral of φ with respect to the Radon measure $\overline{\mathcal{Q}}^{(N)}(t)$ and $\overline{\mathcal{A}}^{(N)}(t)$, respectively, equal $Q_{\varphi}^{(N)}(t)$ and $\overline{A}_{\varphi}^{(N)}(t)$, thus $\{\overline{\mathcal{Q}}^{(N)}(t), t \in [0, \infty)\}$ and $\{\overline{\mathcal{A}}^{(N)}(t), t \in [0, \infty)\}$ can be viewed as $\mathcal{M}_F([0, M) \times \mathbb{R}_+)$-valued càdlàg processes. We now show that the sequences of measure-valued processes $\{\overline{\mathcal{Q}}^{(N)}\}$ and $\{\overline{\mathcal{A}}^{(N)}\}$ are relatively compact.

LEMMA 5.13. *Suppose Assumption 1 is satisfied. Then the sequences* $\{\overline{\mathcal{Q}}^{(N)}\}$ *and* $\{\overline{\mathcal{A}}^{(N)}\}$ *are relatively compact in* $\mathcal{D}_{\mathcal{M}_F([0,M)\times\mathbb{R}_+)}[0, \infty)$.

PROOF. Due to Remark 5.11 and the fact that for $t \geq 0$, the integrals of φ with respect to $\overline{\mathcal{Q}}^{(N)}(t)$ and $\overline{\mathcal{A}}^{(N)}(t)$, respectively, are given by $\overline{Q}_{\varphi}^{(N)}(t)$ and $\overline{A}_{\varphi}^{(N)}(t)$. Lemma 5.10 implies that $\{\overline{\mathcal{Q}}^{(N)}\}$ and $\{\overline{\mathcal{A}}^{(N)}\}$ satisfy Jakubowski's criterion J2. Thus it suffices to verify Jakubowski's J1 criterion for these sequences. Fix $T < \infty$. For $\eta > 0$, define

$$B(\eta) \doteq \frac{2}{\eta} \sup_N \mathbb{E}\big[\overline{A}_{\mathbf{1}}^{(N)}(T)\big], \tag{5.41}$$

which is finite by the first assertion of Lemma 5.6, and let $\{m(n, \eta)\}_{n\in\mathbb{N}} \subset [0, M)$ be a sequence such that $m(n, \eta) \to M$ as $n \to \infty$ and

$$\sup_N \mathbb{E}\big[\overline{A}_{\mathbb{1}_{(m(n,\eta),M)}}^{(N)}(T)\big] \leq \frac{\eta}{n2^{n+1}}, \qquad n \in \mathbb{N}. \tag{5.42}$$

Such a sequence exists by Lemma 5.8(1). Also, define

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

Since $\sup_{\mu\in\mathcal{K}_\eta} \mu([0, M) \times \mathbb{R}_+) \leq B(\eta)$ and

$$\inf_{\substack{C\subset[0,M)\times\mathbb{R}_+: \\ C \text{ compact}}} \sup_{\mu\in\mathcal{K}_\eta} \mu(C^c) \leq \inf_n \sup_{\mu\in\mathcal{K}_\eta} \mu\big((m(n, \eta), M) \times \mathbb{R}_+\big) = 0,$$

$\mathcal{K}_\eta$ is a compact subset of $\mathcal{M}_F([0,M)\times\mathbb{R}_+)$. Moreover, for $\eta>0$ and $N\in\mathbb{N}$, by the monotonicity of $\overline{A}_\varphi^{(N)}(t)$ in t for nonnegative φ, Markov's inequality, (5.41) and (5.42), we have

$$\begin{aligned}
&\mathbb{P}\big(\overline{\mathcal{A}}^{(N)}(t)\notin\mathcal{K}_\eta \text{ for some } t\in[0,T]\big)\\
&\quad\le \mathbb{P}\big(\overline{A}_{\mathbf{1}}^{(N)}(T)\ge B(\eta)\big)+\sum_{n\in\mathbb{N}}\mathbb{P}\bigg(\overline{A}^{(N)}_{\mathbb{1}_{[m(n,\eta),M)}}(T)\ge\frac{1}{n}\bigg)\\
&\quad\le \sup_{N\in\mathbb{N}}\frac{\mathbb{E}[\overline{A}_{\mathbf{1}}^{(N)}(T)]}{B(\eta)}+\sum_{n\in\mathbb{N}} n\sup_{N\in\mathbb{N}}\mathbb{E}\big[\overline{A}^{(N)}_{\mathbb{1}_{[m(n,\eta),M)}}(T)\big]\\
&\quad\le \eta,
\end{aligned}$$

which proves the compact containment condition for $\{\overline{\mathcal{A}}^{(N)}\}$. Due to Corollary 5.5, an exactly analogous argument shows that $\{\overline{\mathcal{Q}}^{(N)}\}$ also satisfies this condition, thus completing the proof of the lemma. □

We are now ready to state the main relative compactness result. Consider the space

$$\mathcal{Y}\doteq\mathbb{R}_+\times(\mathcal{D}_{\mathbb{R}}[0,\infty))^2\times\mathcal{M}_F[0,M)\times\mathcal{D}_{\mathcal{M}_F[0,M)}\times\big(\mathcal{D}_{\mathcal{M}_F([0,M)\times\mathbb{R}_+)}\big)^2$$

equipped with the product metric, and let

$$\overline{Y}^{(N)}\doteq\big(\overline{X}^{(N)}(0),\overline{E}^{(N)},\overline{X}^{(N)},\overline{\nu}_0^{(N)},\overline{\nu}^{(N)},\overline{\mathcal{Q}}^{(N)},\overline{\mathcal{A}}^{(N)}\big),\qquad N\in\mathbb{N}. \tag{5.43}$$

Then $\mathcal{Y}$ is clearly a Polish space, and so, combining Assumption 1 with Lemmas 5.10, 5.12 and 5.13, we arrive at the main result of this section.

THEOREM 5.14. *Suppose Assumption 1 is satisfied. Then the sequence* $\{\overline{Y}^{(N)}\}$ *is relatively compact in the Polish space* $\mathcal{Y}$.

5.4. *Characterization of subsequential limits.* The main result of this section is the following theorem.

THEOREM 5.15. *Suppose Assumptions 1 and 2 are satisfied. Then the limit* $(\overline{X},\overline{\nu})$ *of any subsequence of* $\{(\overline{X}^{(N)},\overline{\nu}^{(N)})\}$ *solves the fluid equations.*

The rest of the section is devoted to the proof of this theorem. Let $(\overline{E},\overline{X}(0),\overline{\nu}_0)$ be the $\mathcal{S}_0$-valued random variable that satisfies Assumption 1, and let $\{\overline{Y}^{(N)}\}$ be the sequence of processes defined in (5.43). Then, by Assumption 1, Theorem 5.14 and the fact that $\overline{M}^{(N)}=\overline{\mathcal{Q}}^{(N)}-\overline{A}^{(N)}\Rightarrow 0$ by Lemma 5.9, there exist $\overline{X}\in\mathcal{D}_{\mathbb{R}_+}[0,\infty)$, $\overline{\nu}\in\mathcal{D}_{\mathcal{M}_F[0,M)}[0,\infty)$ and $\overline{\mathcal{A}}\in\mathcal{D}_{\mathcal{M}_F([0,M)\times\mathbb{R}_+)}[0,\infty)$ such that $\overline{Y}^{(N)}$ converges weakly (along a suitable subsequence) to $\overline{Y}\in\mathcal{Y}$ that has the form

$$\overline{Y}\doteq(\overline{X}(0),\overline{E},\overline{X},\overline{\nu}_0,\overline{\nu},\overline{\mathcal{A}},\overline{\mathcal{A}}).$$

Denoting this subsequence again by $\{\overline{Y}^{(N)}\}$ and invoking the Skorokhod representation theorem, with a slight abuse of notation, we can assume that $\overline{Y}^{(N)} \to \overline{Y}$ $\mathbb{P}$-almost surely.

We now identify some properties of the limit that will be used to prove Theorem 5.15. First, note that we immediately have $\langle \mathbf{1}, \overline{\nu}^{(N)} \rangle \to \langle \mathbf{1}, \overline{\nu} \rangle$ almost surely and $\overline{D}^{(N)} = \overline{Q}_{\mathbf{1}}^{(N)} \to \overline{A}_{\mathbf{1}}$, where $\overline{Q}_{\mathbf{1}}^{(N)}$ and $\overline{A}_{\mathbf{1}}$ are defined in (5.1) and (5.23) with $\varphi = \mathbf{1}$. When combined with (2.5) and (2.6), this implies that

$$\overline{X} = \overline{X}(0) + \overline{E} - \overline{Q}_{\mathbf{1}} = \overline{X}(0) + \overline{E} - \overline{A}_{\mathbf{1}}, \tag{5.44}$$

and that $\overline{K}^{(N)}$ converges a.s. to $\overline{K}$, where

$$\overline{K}(t) \doteq \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \overline{A}_{\mathbf{1}}, \qquad t \in [0, \infty). \tag{5.45}$$

Moreover, from (2.10) it follows that nonidling condition (3.7) holds. Comparing (5.44) and (5.45) with (3.6) and (3.8), it is clear that in order to prove that $(\overline{X}, \overline{\nu})$ satisfies the fluid equations, it is necessary to show that $\overline{Q}_{\mathbf{1}} = \overline{A}_{\mathbf{1}} = \overline{D}$, where $\overline{D}$ is defined in terms of $\overline{\nu}$ via (3.9). If h is continuous and uniformly bounded on $[0, M)$ (as is the case, e.g., when G is a lognormal distribution), then this is a simple consequence of the (almost sure) weak convergence of $\overline{\nu}^{(N)}$ to $\overline{\nu}$ and the definition of $A_{\mathbf{1}}^{(N)}$ given in (5.23). However, as we show below, some additional work is required to justify this convergence for general h (that satisfies Assumption 2). We start in Lemma 5.16 by establishing the bound (5.46) for certain integrals with respect to $\overline{\nu}$. This bound is then used in Proposition 5.17 to show that, under mild additional conditions on h, $\overline{A}_{\mathbf{1}}$ equals $\overline{D}$. Next, Lemma 5.18 establishes sufficient conditions under which $\overline{\nu}_s$ is absolutely continuous with respect to Lebesgue measure on $[0, M)$, for every s. All these results are then combined to complete the proof of Theorem 5.15 at the end of the section.

LEMMA 5.16. *For $m \in [0, M)$ and every $\ell \in \mathcal{L}^1_{\mathrm{loc}}[0, M)$ with support in $[0, m]$, there exists $\tilde{L}(m, T) < \infty$ such that*

$$\left| \int_0^\infty \langle \ell, \overline{\nu}_s \rangle \, ds \right| \le \tilde{L}(m, T) \int_{[0, M)} |\ell(x)| \, dx. \tag{5.46}$$

PROOF. To establish the lemma, it suffices to show that for every $m \in [0, M)$, $T \in (0, \infty)$ and $\varphi \in \mathcal{C}_c^{1,1}([0, M) \times \mathbb{R}_+)$ with $\mathrm{supp}(\varphi) \subset [0, m] \times [0, T]$, there exists $C(m, T) < \infty$ such that

$$\left| \int_0^T \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}_s \rangle \, ds \right| \le \|\varphi\|_\infty C(m, T). \tag{5.47}$$

Then the measure $\tilde{\gamma}$ on $\mathbb{R}^2$, defined by

$$\iint_{\mathbb{R}^2} \varphi(x, s) \tilde{\gamma}(dx, ds) \doteq - \int_0^\infty \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}_s \rangle \, ds,$$

$$\varphi \in \mathcal{C}_c^{1,1}\big([0, M) \times \mathbb{R}_+\big),$$

is a Radon measure on $\mathbb{R}^2$ with support in $[0, M) \times \mathbb{R}_+$ and, moreover, $\{\overline{\nu}_s\}_{s\geq 0}$ satisfies the simplified age equation for $\tilde{\gamma}$. Therefore, (5.46) follows from Proposition 4.15. To establish (5.47), fix m, T and φ as above. For any $\varepsilon > 0$, substituting $t = T + \varepsilon$ in (5.4), the term $\langle \varphi(\cdot, T+\varepsilon), \overline{\nu}^{(N)}_{T+\varepsilon}\rangle$ equals zero. Therefore, rearranging the remaining terms, we obtain for every $N \in \mathbb{N}$,

$$\begin{aligned}\int_0^{T+\varepsilon} &\langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}^{(N)}_s\rangle\, ds \\ &= \langle \varphi(\cdot, 0), \overline{\nu}^{(N)}_0\rangle + \int_{[0,T+\varepsilon]} \varphi(0, u)\, d\overline{K}^{(N)}_u \\ &\quad + \overline{M}^{(N)}_\varphi(T+\varepsilon) + \overline{A}^{(N)}_\varphi(T+\varepsilon).\end{aligned}$$

Sending $\varepsilon \to 0$ and using the right-continuity of the processes, the fact that $\overline{K}^{(N)}$ is nondecreasing and the bound (5.25) on $|\overline{A}^{(N)}_\varphi(T)|$, this implies that

$$\left|\int_0^T \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}^{(N)}_s\rangle ds\right| \leq \|\varphi\|_\infty C^{(N)}(m, T) + \overline{M}^{(N)}_\varphi(T), \tag{5.48}$$

where

$$\begin{aligned}C^{(N)}(m, T) \doteq{}& \langle \mathbf{1}, \overline{\nu}^{(N)}_0\rangle + \overline{K}^{(N)}(T) \\ &+ \big(\overline{X}^{(N)}(0) + \overline{E}^{(N)}(T)\big) \int_0^m h(x)\, dx.\end{aligned}$$

Due to (2.5), (2.6) and the limits $\overline{E}^{(N)} \to \overline{E}$ and $\overline{X}^{(N)}(0) \to \overline{X}(0)$ as $N \to \infty$, it follows that

$$\limsup_{N\to\infty} C^{(N)}(m, T) \leq C(m, T) \doteq 2\big(1 + \overline{X}(0) + \overline{E}(T)\big)\left[1 \vee \left(\int_0^m h(x)\, dx\right)\right].$$

Taking limits as $N \to \infty$ on both sides of (5.48), recalling that $\overline{M}^{(N)}_\varphi(T) \to 0$ (see Lemma 5.9) and observing that the left-hand side converges to $|\int_0^T \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}_s\rangle\, ds|$ due to the bounded convergence theorem, the fact that $\varphi_x(\cdot, s) + \varphi_s(\cdot, s) \in \mathcal{C}_c([0, M) \times \mathbb{R}_+)$ and $\overline{\nu}^{(N)}_s \xrightarrow{w} \overline{\nu}_s$ for a.e. $s \in [0, T]$, we obtain (5.47). □

PROPOSITION 5.17. *If h satisfies Assumption 2, then for every $\varphi \in \mathcal{C}_b([0, M) \times \mathbb{R}_+)$,*

$$\overline{A}_\varphi(t) = \int_0^t \langle \varphi(\cdot, s)h(\cdot), \overline{\nu}_s\rangle\, ds, \qquad t \in [0, \infty). \tag{5.49}$$

In particular, $\overline{A}_\varphi$ is absolutely continuous and $\overline{A}_{\mathbf{1}} = \overline{D}$.

PROOF. We will show (5.49) only for the case when $\varphi = \mathbf{1}$ because the proof for arbitrary $\varphi \in \mathcal{C}_b([0, M) \times \mathbb{R}_+)$ is exactly analogous. Note that for any $m \in [0, M)$, we have the elementary bound

$$\sup_{t\in[0,T]} \left|\overline{A}^{(N)}_{\mathbf{1}}(t) - \overline{D}(t)\right| \le R^{(N)}_1(m) + R^{(N)}_2(m) + R_3(m), \tag{5.50}$$

where, setting $h_m \doteq h\mathbb{1}_{[0,m]}$, we define

$$\begin{aligned}
R^{(N)}_1(m) &\doteq \int_0^T \left(\int_{(m,M)} h(x)\overline{\nu}^{(N)}_s(dx)\right) ds,\\
R^{(N)}_2(m) &\doteq \int_0^T \left|\left(\int_{[0,M)} h_m(x)\overline{\nu}^{(N)}_s(dx)\right) - \left(\int_{[0,M)} h_m(x)\overline{\nu}_s(dx)\right)\right| ds,\\
R_3(m) &\doteq \int_0^T \left(\int_{(m,M)} h(x)\overline{\nu}_s(dx)\right) ds.
\end{aligned}$$

We now analyze each of the above terms. First, by Lemma 5.8(1), it follows that

$$\lim_{m\to M} \sup_N \mathbb{E}\big[R^{(N)}_1(m)\big] = 0. \tag{5.51}$$

Next, note that h_m is with compact support and integrable on $[0, M)$. Hence, there exists a sequence $\{h_{m,k}\}_{k\in\mathbb{N}} \subset \mathcal{C}_c[0, M)$ that, as $k \to \infty$, converges in $\mathcal{L}^1[0, M)$ to h_m. For every $k, m \in \mathbb{N}$, we have the bound

$$\begin{aligned}
R^{(N)}_2(m) \le{}& \int_0^T \left(\int_{[0,M)} |h_m(x) - h_{m,k}(x)|\overline{\nu}^{(N)}_s(dx)\right) ds\\
&+ \int_0^T \left|\left(\int_{[0,M)} h_{m,k}(x)\overline{\nu}^{(N)}_s(dx)\right) - \left(\int_{[0,M)} h_{m,k}(x)\overline{\nu}_s(dx)\right)\right| ds\\
&+ \int_0^T \left(\int_{[0,M)} |h_m(x) - h_{m,k}(x)|\overline{\nu}_s(dx)\right) ds.
\end{aligned} \tag{5.52}$$

Applying Proposition 5.7, with $\ell = |h_m - h_{m,k}|$, $\varphi = \mathbf{1}$, $r = 0$ and $t = T$, and taking first expectations and then the supremum over N, we see that

$$\begin{aligned}
&\sup_N \mathbb{E}\left[\int_0^T \left(\int_{[0,M)} |h_{m,k}(x) - h_m(x)|\overline{\nu}^{(N)}(dx)\right) ds\right]\\
&\qquad \le C(T)\int_0^M |h_{m,k}(x) - h_m(x)|\, dx,
\end{aligned}$$

where $C(T) \doteq 1 + \sup_N \mathbb{E}[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(T)]$ is finite due to properties 1 and 2 of Assumption 1. Taking limits as $k \to \infty$ and using the $\mathcal{L}^1$-convergence of $h_{m,k}$ to h_m, we then obtain

$$\lim_{k\to\infty} \sup_N \mathbb{E}\left[\int_0^T \left(\int_{[0,M)} |h_m(x) - h_{m,k}(x)|\overline{\nu}^{(N)}_s(dx)\right) ds\right] = 0. \tag{5.53}$$

Likewise, an application of Lemma 5.16 with $\ell = |h_{m,k} - h_m|$ yields

$$\begin{aligned}\lim_{k\to\infty} \mathbb{E}\Big[\int_0^T \Big(\int_{[0,M)} |h_m(x) - h_{m,k}(x)|\overline{\nu}_s(dx)\Big)\,ds\Big] \\ \le \lim_{k\to\infty} \tilde{L}(m,T)\Big(\int_0^M |h_m(x) - h_{m,k}(x)|\,dx\Big) = 0.\end{aligned} \tag{5.54}$$

Moreover, by the weak convergence $\overline{\nu}_s^{(N)} \xrightarrow{w} \overline{\nu}_s$ as $N \to \infty$ for a.e. $s \in [0,T]$, the fact that $h_{m,k}$ is bounded and continuous and by the bounded convergence theorem, we see that

$$\begin{aligned}\limsup_{N\to\infty} \mathbb{E}\Big[\int_0^T \Big|\Big(\int_{[0,M)} h_{m,k}(x)\overline{\nu}_s^{(N)}(dx)\Big) \\ - \Big(\int_{[0,M)} h_{m,k}(x)\overline{\nu}_s(dx)\Big)\Big|\,ds\Big] = 0.\end{aligned} \tag{5.55}$$

Taking first the expectation on both sides of (5.52), next the limit superior over N and then limits as $k \to \infty$ of the right-hand side, and using (5.53)–(5.55), we obtain

$$\limsup_{N\to\infty} \mathbb{E}\big[R_2^{(N)}(m)\big] = 0. \tag{5.56}$$

We now consider the third term, using Assumption 2. If h is bounded (say by B) on (m_0, M), then by the bounded convergence theorem,

$$\lim_{m\to M} \mathbb{E}[R_3(m)] \le \lim_{m\to M} B\int_0^T \overline{\nu}_s(m,M)\,ds = B\int_0^T \lim_{m\to M} \overline{\nu}_s(m,M)\,ds = 0.$$

On the other hand, suppose h is lower semicontinuous on (m_0, M). Then for any $m \ge m_0$, $h\mathbb{1}_{(m,M)}$ is lower semicontinuous on $[0, M)$ since it is identically zero on $[0, m)$, coincides with the lower semicontinuous function h on (m, ∞) and at m, the nonnegativity of h implies $\mathbb{1}_{(m,M)}(m)h(m) = 0 \le \lim_{x\to m} \mathbb{1}_{(m,M)}(x)$. Together with Theorem A.3.12 of [4] and the fact that $\mathbb{P}$-a.s., $\overline{\nu}_s^{(N)} \xrightarrow{w} \overline{\nu}_s$ as $N \to \infty$ for a.e. $s \in [0,T]$, this implies that for any such s and $m \ge m_0$,

$$\int_{[0,M)} \mathbb{1}_{(m,M)}(x)h(x)\overline{\nu}_s(dx) \le \liminf_{N\to\infty} \int_{[0,M)} \mathbb{1}_{(m,M)}(x)h(x)\overline{\nu}_s^{(N)}(dx).$$

Integrating both sides over $s \in [0,T]$ and taking expectations, an application of Fatou's lemma yields

$$\mathbb{E}[R_3(m)] \le \liminf_{N\to\infty} \mathbb{E}\big[R_1^{(N)}(m)\big].$$

Taking the limit as $m \uparrow M$ and invoking (5.51) we conclude that in this case as well, we have

$$\limsup_{m\to M} \mathbb{E}[R_3(m)] = 0. \tag{5.57}$$

Taking first expectations and then the limit superior, as $N \to \infty$, of both sides of (5.50), an application of Fatou's lemma and the limit $\overline{A}_{\mathbf{1}}^{(N)} \to \overline{A}_{\mathbf{1}}$ as $N \to \infty$, show that for every $m \in [0, M)$,

$$\begin{aligned}\mathbb{E}\Big[\sup_{t\in[0,T]} |\overline{A}_{\mathbf{1}}(t) - \overline{D}(t)|\Big] &\le \limsup_{N\to\infty} \mathbb{E}\Big[\sup_{t\in[0,T]} |\overline{A}_{\mathbf{1}}^{(N)}(t) - \overline{D}(t)|\Big] \\ &\le \limsup_{N\to\infty} \mathbb{E}\big[R_1^{(N)}(m) + R_2^{(N)}(m)\big] + \mathbb{E}[R_3(m)].\end{aligned}$$

Sending $m \to M$ on the right-hand side and invoking (5.51), (5.56) and (5.57), we see that the right-hand side converges to zero. This shows that $\sup_{t\in[0,T]}|\overline{A}_{\mathbf{1}}(t) - \overline{D}(t)| = 0$ a.s. Because T is arbitrary, this proves the first assertion of the proposition. The second assertion is an immediate consequence of the first. □

LEMMA 5.18. *If $\overline{\nu}_0$ and $\overline{E}$ are absolutely continuous, then $\overline{\nu}_s$ is also absolutely continuous for every $s \in [0, \infty)$.*

PROOF. For simplicity, we will assume that $M = \infty$ for the rest of the proof. The case $M < \infty$ is analogous and is left to the reader. Since $\overline{E}$ is absolutely continuous by assumption and $\overline{A}_{\mathbf{1}}$ is absolutely continuous by Proposition 5.17, (5.44) allows us to deduce that $\overline{X}$ is absolutely continuous. The nonidling condition (3.7) and the mass balance equation (3.8) then show that $\langle \mathbf{1}, \overline{\nu}\rangle$ and $\overline{K}$ are also absolutely continuous. Fix $T < \infty$, and let C_ε denote the set of collections of finite disjoint intervals $(a_i, b_i) \subset [0, T]$, $i = 1, \dots, n$, such that $\sum_{i=1}^n (b_i - a_i) \le \varepsilon$. Given $\delta > 0$, choose $\varepsilon > 0$ small enough so that

$$\sup_{\{(a_i,b_i)\}\in C_\varepsilon} \left[\overline{\nu}_0\left(\bigcup_{i=1}^n [a_i, b_i]\right)\right] \vee \left[\sup_{s\in[0,T]} \sum_{i=1}^n \overline{K}(s - b_i, s - a_i)\right] < \frac{\delta}{2},$$

where we recall that $\overline{K}(s,t] = \overline{K}(t) - \overline{K}(s)$, and when $t < 0$, $\overline{K}(s,t]$ is defined to be zero. Now, consider a particular collection $\{(a_i, b_i)\} \in C_\varepsilon$. Fix $s \in [0, T]$. Define $J_1 = \{i \in \{1, \dots, n\} : a_i - s \ge 0\}$ and let $J_2 = \{1, \dots, n\} \setminus J_1$. Then we have

$$\begin{aligned}\sum_{i=1}^n \overline{\nu}_s^{(N)}(a_i, b_i) &= \sum_{i\in J_1} \overline{\nu}_s^{(N)}(a_i, b_i) + \sum_{i\in J_2} \overline{\nu}_s^{(N)}(a_i, b_i) \\ &\le \sum_{i\in J_1} \overline{\nu}_0^{(N)}(a_i - s, b_i - s) + \sum_{i\in J_2} \overline{\nu}_{s-a_i}^{(N)}(0, b_i - a_i) \\ &\le \sum_{i\in J_1} \overline{\nu}_0^{(N)}(a_i - s, b_i - s) \\ &\quad + \sum_{i\in J_2} \overline{K}^{(N)}(s - b_i, s - a_i).\end{aligned}$$

Taking limits as $N \to \infty$ and using the fact that $\overline{K}^{(N)} \to \overline{K}$, Assumption 1(3) and the Portmanteau theorem, we see that

$$\begin{aligned}\overline{\nu}_s\Bigg(\bigcup_{i=1}^{n}(a_i,b_i)\Bigg) &\le \liminf_N \overline{\nu}_s^{(N)}\Bigg(\bigcup_{i=1}^{n}(a_i,b_i)\Bigg) = \liminf_N \sum_{i=1}^{n}\overline{\nu}_s^{(N)}(a_i,b_i)\\ &\le \limsup_N \sum_{i\in J_1}\overline{\nu}_0^{(N)}(a_i-s,b_i-s) + \sum_{i\in J_2}\lim_N \overline{K}^{(N)}(s-b_i,s-a_i)\\ &\le \overline{\nu}_0\Bigg(\bigcup_{i\in J_1}[a_i-s,b_i-s]\Bigg) + \sum_{i\in J_2}\overline{K}(s-b_i,s-a_i).\end{aligned}$$

Taking the supremum over all collections of intervals in C_ε, we conclude that for every $\delta > 0$, there exists $\varepsilon > 0$ such that

$$\sup_{\{(a_i,b_i)\}\in C_\varepsilon}\sum_{i=1}^{n}\overline{\nu}_s(a_i,b_i) \le \delta,$$

which shows that $\overline{\nu}_s$ is absolutely continuous. □

Combining the above results, we now have a proof of the main result of this section.

PROOF OF THEOREM 5.15. Let $\tilde{\Omega}$ be the set of full $\mathbb{P}$-measure on which the properties stated in Assumptions 1 and 2 hold, and $\overline{Y}^{(N)}(\omega) \to \overline{Y}(\omega)$ for all $\omega \in \tilde{\Omega}$. Fix $\omega \in \tilde{\Omega}$ (and suppress it from the notation), and then choose $t \in [0,\infty)$ such that at that ω, $\overline{\nu}_t^{(N)} \xrightarrow{w} \overline{\nu}_t$, $\overline{A}^{(N)}(t) \xrightarrow{w} \overline{A}(t)$, $\overline{Q}^{(N)}(t) \xrightarrow{w} \overline{A}(t)$, $\overline{E}^{(N)}(t) \to \overline{E}(t)$ and $\overline{X}^{(N)}(t) \to \overline{X}(t)$ as $N \to \infty$. Note that this occurs for t outside a countable set of times (which may depend on ω). Moreover, because $\overline{K}^{(N)}(t)$ can be written in terms of $\overline{E}^{(N)}(t)$, $\overline{X}^{(N)}(t)$, $\langle \mathbf{1}, \overline{\nu}_t^{(N)}\rangle$ and the initial conditions due to (2.5) and (2.6), and, likewise, $\overline{K}(t)$ can be expressed in terms of the $\overline{E}(t)$, $\overline{X}(t)$, $\langle \mathbf{1}, \overline{\nu}_t\rangle$ and limits of the initial conditions using (3.6) and (3.8), it follows from the limits at t assumed above and Assumption 1 that $\overline{K}^{(N)}(t) \to \overline{K}(t)$.

Now, fix $\varphi \in \mathcal{C}_c^{1,1}([0,M)\times\mathbb{R}_+)$. By Proposition 5.17, as $N \to \infty$, we have

$$\overline{Q}_\varphi^{(N)}(t) \to \overline{A}_\varphi(t) = \int_0^t \langle \varphi(\cdot,s)h(\cdot,s), \overline{\nu}_s\rangle\, ds. \tag{5.58}$$

In particular, this implies $\overline{D}^{(N)}(t) = \overline{Q}_{\mathbf{1}}^{(N)}(t) \to \overline{D}(t)$ and $\overline{D}(t) < \infty$ for every $t \in [0,\infty)$, which shows that condition (3.4) of the fluid equations is satisfied. Also, when combined with (5.44), (5.45) and the fact that the nonidling condition is satisfied (see the discussion immediately after the statement of Theorem 5.15), this implies that the fluid equations (3.6)–(3.8) are also satisfied.

It only remains to show that $\overline{\nu}$ and $\overline{K}$ satisfy (3.5) for φ. First, dividing (5.4) by N, we obtain

$$\langle \varphi(\cdot,t), \overline{\nu}_t^{(N)} \rangle = \langle \varphi(\cdot,0), \overline{\nu}_0^{(N)} \rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \overline{\nu}_s^{(N)} \rangle \, ds - \overline{Q}_\varphi^{(N)}(t) + \int_{[0,t]} \varphi(0,u) \, d\overline{K}^{(N)}(u).$$

By the choice of ω and t, as $N \to \infty$, $\overline{\nu}_0^{(N)} \xrightarrow{w} \overline{\nu}_0$, $\overline{\nu}_s^{(N)} \xrightarrow{w} \overline{\nu}_s$ for a.e. $s \in [0,t]$, $\overline{\nu}_t^{(N)} \xrightarrow{w} \overline{\nu}_t$. Due to the boundedness of $\varphi(\cdot,t)$ and $\varphi_x(\cdot,s) + \varphi_s(\cdot,s)$, $s \in [0,t]$, the bounded convergence theorem then implies that, as $N \to \infty$,

$$\langle \varphi(\cdot,0), \overline{\nu}_0^{(N)} \rangle \to \langle \varphi(\cdot,0), \overline{\nu}_0 \rangle, \qquad \langle \varphi(\cdot,t), \overline{\nu}_t^{(N)} \rangle \to \langle \varphi(\cdot,t), \overline{\nu}_t \rangle$$

and

$$\int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \overline{\nu}_s^{(N)} \rangle \, ds \to \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \overline{\nu}_s \rangle \, ds.$$

On the other hand, because $\overline{K}^{(N)}(s) \to \overline{K}(s)$ for all continuity points s of $\overline{K}$, the associated sequence of Stieltjes measures $d\overline{K}^{(N)}$ converges vaguely to the corresponding Stieltjes measure $d\overline{K}$, as $N \to \infty$. Since $\overline{K}^{(N)}(t) \to \overline{K}(t)$ and $\varphi(0,\cdot) \in \mathcal{C}_c(\mathbb{R}_+)$, this implies that, as $N \to \infty$,

$$\lim_{n\to\infty} \int_{[0,t]} \varphi(0,u) \, d\overline{K}^{(N)}(u) = \int_{[0,t]} \varphi(0,u) \, d\overline{K}(u).$$

Combining the last four displays with (5.58), it follows that the fluid equation (3.5) is satisfied for all but countably many t. By right-continuity (with respect to t) of each of the terms in (3.5), we conclude that (3.5) is satisfied for all $t \in [0,\infty)$. This completes the proof of the desired result that a.s., $(\overline{X}, \overline{\nu})$ satisfies the fluid equations. □

We now obtain Theorem 3.7 as an immediate corollary.

PROOF OF THEOREM 3.7. The first statement of Theorem 3.7 is a direct consequence of Theorem 5.14. The remainder of the theorem follows from Theorem 5.15, the uniqueness of solutions to the fluid equations established in Theorem 3.5 and the usual standard argument by contradiction that shows that the original sequence converges to the solution of the fluid limit whenever the latter is unique. □

6. Convergence of the fluid limit to equilibrium. Throughout this section we assume that h satisfies Assumption 2 and that the initial condition $(\overline{E}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$ is such that $\overline{E}$ is absolutely continuous with derivative denoted

by $\overline{\lambda}(\cdot)$, that is,

$$\overline{E}(t) = \int_0^t \overline{\lambda}(s)\,ds, \qquad t \in [0,\infty).$$

By Theorems 3.5 and 3.7, there exists a unique solution $(\overline{X},\overline{\nu})$ to the associated fluid equations, which is characterized by (3.6), (3.11) and (3.12). The goal of this section is to prove Theorem 3.9, which describes the large-time behavior of $(\overline{X},\overline{\nu})$. First, in Proposition 6.1, the case when the system starts empty is analyzed and, in addition, a certain comparison result is established between such systems and systems with more general initial conditions that have the same (fluid) arrival rate $\overline{\lambda}(\cdot)$. The proof of Theorem 3.9 is provided at the end of the section, building on two preliminary results obtained in Lemmas 6.2 and 6.3.

Recall from (3.13) that $\overline{\nu}_*$ is the probability measure with density $1-G$, and also recall the definition of monotonic weak convergence stated before Theorem 3.9.

PROPOSITION 6.1. *Let $(\overline{X},\overline{\nu})$ be the unique solution to the fluid equations associated with the initial condition $(\overline{E},0,\tilde{\mathbf{0}}) \in \mathcal{S}_0$. If*

$$\tau_1 \doteq \inf\Big\{t>0 : \int_0^t \big(1-G(t-s)\big)\overline{\lambda}(s)\,ds = 1\Big\}, \tag{6.1}$$

then we have the following properties:

(1) *For $t \in [0,\tau_1)$,*

$$\langle \mathbf{1}, \overline{\nu}_t\rangle = \overline{X}(t) = \int_0^t \big(1-G(t-s)\big)\overline{\lambda}(s)\,ds$$

and, as $t \to \tau_1$, for every function $f \in \mathcal{C}_b(\mathbb{R}_+)$,

$$\langle f, \overline{\nu}_t\rangle \to \int_0^{\tau_1} f(t-s)\big(1-G(t-s)\big)\overline{\lambda}(s)\,ds.$$

(2) *Suppose $\overline{\lambda}(\cdot)$ is a constant equal to $\overline{\lambda} \in [0,1]$. Then, as $t\to\infty$, $\overline{X}(t) = \langle \mathbf{1}, \overline{\nu}_t\rangle \to \overline{\lambda}$ monotonically and $\overline{\nu}_t$ converges weakly monotonically up to $\overline{\lambda}\overline{\nu}_*$.*

(3) *Suppose $(\overline{X}^\diamond,\overline{\nu}^\diamond)$ is the unique solution to the fluid equations associated with any other initial condition $(\overline{E},\overline{X}^\diamond(0),\overline{\nu}_0^\diamond)$ that has the same fluid cumulative arrival process. Then*

$$\langle \mathbf{1}, \overline{\nu}_t^\diamond\rangle \ge \langle \mathbf{1}, \overline{\nu}_t\rangle \qquad \text{for all } t \in [0,\tau_1). \tag{6.2}$$

Moreover, if $\overline{\lambda}(\cdot)$ is a constant equal to $\overline{\lambda} \in [0,1]$, then

$$\langle \mathbf{1}, \overline{\nu}_t^\diamond\rangle \ge \overline{\lambda}\int_0^t \big(1-G(r)\big)\,dr \qquad \text{for all } t \in [0,\tau_1). \tag{6.3}$$

PROOF. Let $(\overline{X},\overline{\nu})$ be the unique solution to the fluid equations associated with the initial conditions $(\overline{E},0,\tilde{\mathbf{0}})$, and define

$$\tau \doteq \inf\{t>0 : \langle \mathbf{1}, \overline{\nu}_t\rangle = 1\},$$

with the usual convention that $\inf \varnothing = \infty$. Then $\tau > 0$ by the right-continuity of $t \mapsto \langle \mathbf{1}, \overline{\nu}_t \rangle$ and for $t \in (0, \tau)$, $\langle \mathbf{1}, \overline{\nu}_t \rangle < 1$. The nonidling property in (3.7) then implies that $\overline{X}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle < 1$ for $t \in [0, \tau)$. Combining relations (3.6) and (3.8), this in turn implies that $\overline{K}(t) = \overline{E}(t)$ for $t \in [0, \tau)$. Since, by (3.11), $\overline{\nu}$ satisfies the age equation associated with $\overline{\nu}_0 \equiv \tilde{\mathbf{0}}$ and $\overline{K}$, we have for any $f \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \in [0, \tau)$,

$$\begin{aligned}\langle f, \overline{\nu}_t \rangle &= \int_{[0,t]} f(t-s)\big(1 - G(t-s)\big)\, d\overline{E}(s) \\ &= \int_0^t f(t-s)\big(1 - G(t-s)\big)\overline{\lambda}(s)\, ds.\end{aligned}$$

Substituting $f = \mathbf{1}$, the left-hand side above is equal to $\overline{X}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle$, from which it is easy to see that $\tau = \tau_1$ since the integrand on the right-hand side is continuous in t. This proves property 1.

Next, consider the time-homogeneous setting where $\overline{\lambda}(\cdot)$ equals a constant $\overline{\lambda} \in [0, 1]$. The case $\overline{\lambda} = 0$ is trivial and when $\overline{\lambda} \in (0, 1]$, we can rewrite

$$\tau_1 = \inf\left\{ t > 0 : \int_0^t \big(1 - G(x)\big)\, dx = 1/\overline{\lambda} \right\}.$$

This shows that $\tau_1 = \infty$ when either $\overline{\lambda} < 1$, or $\overline{\lambda} = 1$ and $M = \infty$, and so, by property 1, it follows that $\overline{\nu}_t \overset{w}{\to} \overline{\lambda}\overline{\nu}_*$ monotonically as $t \to \infty$. Thus property 2 is proved in this case. Now, suppose $\overline{\lambda} = 1$ and $M < \infty$. Then analogous reasoning shows that $\tau_1 = M < \infty$, $\overline{X}(\tau_1) = 1$ and $\overline{\nu}_{\tau_1} = \overline{\nu}_*$. When combined with the nonanticipative property stated in Lemma 3.4 and the fact that $(\mathbf{1}, \overline{\nu}_*)$ is an invariant solution for the fluid equations with initial conditions $(\mathrm{id}, \mathbf{1}, \overline{\nu}_*)$ (see Remark 3.8), this shows that $(\overline{X}(t), \overline{\nu}_t) = (\mathbf{1}, \overline{\nu}_*)$ for all $t \geq \tau_1$ and once again property 2 follows.

We now turn to the last property. We first show that for every $t \in (0, \tau_1)$, $\overline{X}^\diamond(t) \geq \overline{X}(t)$. To this end we combine (3.6), (3.8) and the fact that $\langle \mathbf{1}, \overline{\nu}_t^\diamond \rangle \leq \overline{X}^\diamond(t)$ due to (3.7) to obtain

$$\overline{K}^\diamond(t) = \langle \mathbf{1}, \overline{\nu}_t^\diamond \rangle - \langle \mathbf{1}, \overline{\nu}_0^\diamond \rangle + \overline{X}^\diamond(0) + \overline{E}(t) - \overline{X}^\diamond(t) \leq \overline{X}^\diamond(0) - \langle \mathbf{1}, \overline{\nu}_0^\diamond \rangle + \overline{E}(t).$$

When combined with (3.6), (3.8) and (4.5), this implies

$$\begin{aligned}\overline{X}^\diamond(t) &= \overline{X}^\diamond(0) + \overline{E}(t) - \int_{[0,M)} \frac{G(x+t) - G(x)}{1 - G(x)} \overline{\nu}_0^\diamond(dx) - \int_0^t g(t-s)\overline{K}^\diamond(s)\, ds \\ &\geq \overline{X}^\diamond(0) + \overline{E}(t) - \int_{[0,M)} \frac{G(x+t) - G(x)}{1 - G(x)} \overline{\nu}_0^\diamond(dx) \\ &\quad - \int_0^t g(t-s)\big(\overline{X}^\diamond(0) - \langle \mathbf{1}, \overline{\nu}_0^\diamond \rangle + \overline{E}(s)\big)\, ds \\ &\geq \overline{X}^\diamond(0) + \overline{E}(t) - \langle \mathbf{1}, \overline{\nu}_0^\diamond \rangle - \int_0^t \overline{E}(s) g(t-s)\, ds - \big(\overline{X}^\diamond(0) - \langle \mathbf{1}, \overline{\nu}_0^\diamond \rangle\big) G(t).\end{aligned}$$

Since $G(t) \le 1$ and $\overline{X}^{\diamond}(0) \ge \langle \mathbf{1}, \overline{\nu}_0^{\diamond} \rangle$, we have in fact

$$\overline{X}^{\diamond}(t) \ge \overline{E}(t) - \int_0^t \overline{E}(s) g(t-s)\, ds. \tag{6.4}$$

However, due to (3.12) and the fact that property 1 implies $\overline{X}(s) < 1$ for $s \in [0, \tau_1)$, we know that $\overline{E}(s) = \overline{K}(s)$ for all $s \in [0, \tau_1)$. Together with the relations (3.6), (3.8), (4.5) and the fact that $\overline{X}(0) = 0$ and $\overline{\nu}_0 = \tilde{\mathbf{0}}$, this shows that the right-hand side of (6.4) equals $\overline{X}(t)$. This shows inequality (6.2) holds because if $\overline{X}^{\diamond}(t) \ge 1$, then $\langle \mathbf{1}, \overline{\nu}_t^{\diamond} \rangle = 1 \ge \langle \mathbf{1}, \overline{\nu}_t \rangle$, whereas if $\overline{X}^{\diamond}(t) < 1$, then $\langle \mathbf{1}, \overline{\nu}_t^{\diamond} \rangle = \overline{X}^{\diamond}(t) \ge \overline{X}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle$. Finally, note that when $\overline{\lambda}(\cdot) \equiv \overline{\lambda} \in [0, 1]$, by property 1 we see that for $t \in [0, \tau_1)$, $\langle \mathbf{1}, \overline{\nu}_t \rangle$ equals the right-hand side of (6.3). Thus (6.3) follows immediately from (6.2). □

The result of Theorem 3.9(1) now follows from Proposition 6.1(2). We now turn to the proof of Theorem 3.9(2), which involves a comparison of the measure-valued function $\{\overline{\nu}_s\}$, which is the solution of the age equation corresponding to $\overline{\nu}_0$ and $\overline{K}$, with another "reference" measure-valued function $\{\pi_s\}$ that solves the age equation associated with an initial condition of the form $\pi_0 = \overline{\nu}_T$ for some $T < \infty$, and the function Z defined in (6.5) below. Although we do not explicitly use this below, the reference function $\{\pi_s\}$ has the property that its total mass remains constant for all times or, equivalently, viewing Z as the cumulative entry into service in a reference fluid server system, that the cumulative entry equals the cumulative departures in that system at all times. This makes the long-time behavior of the reference function $\{\pi_s\}$ easier to analyze. First, in Lemma 6.2, it is shown that when $\langle \mathbf{1}, \overline{\nu}_T \rangle = 1$, π_s converges weakly to $\overline{\nu}_*$ as $s \to \infty$. Then, in Lemma 6.3, an estimate is obtained which is used in the proof of Theorem 3.9(2) to show that the difference between the original function $\{\overline{\nu}_s\}$ and the reference function $\{\pi_s\}$ vanishes as $s \to \infty$.

Recall that U is the renewal function associated with the service distribution that has cumulative distribution function G, and let u denote its density, which exists since G has a density (see Proposition 2.7 of Section V in [1]).

LEMMA 6.2. *Given $\pi_0 \in \mathcal{M}_{\le 1}[0, M)$, suppose $Z \in \mathcal{I}_0[0, \infty)$ and $\{\pi_t\} \in \mathcal{D}_{\mathcal{M}_F}[0, \infty)$ are defined as follows*:

$$Z(t) \doteq \int_{[0,t]} \left(\int_{[0,M)} \frac{G(x+t-s) - G(x)}{1 - G(x)} \pi_0(dx) \right) dU(s), \qquad t \in [0, \infty), \tag{6.5}$$

and, for $f \in \mathcal{C}_b[0, M)$ and $t \in [0, \infty)$,

$$\begin{aligned} \langle f, \pi_t \rangle = {} & \int_{[0,M)} f(x+t) \frac{1 - G(x+t)}{1 - G(x)} \pi_0(dx) \\ & + \int_{[0,t]} f(t-s)\big(1 - G(t-s)\big)\, dZ(s). \end{aligned} \tag{6.6}$$

Then for every $f \in \mathcal{C}_b[0, M)$,

$$\lim_{t\to\infty} \langle f, \pi_t\rangle = \langle \mathbf{1}, \pi_0\rangle \langle f, \overline{\nu}_*\rangle. \tag{6.7}$$

PROOF. It is easy to see that Z is absolutely continuous with density

$$\frac{dZ}{dt}(t) = \int_0^t \Big(\int_{[0,M)} \frac{g(x+t-s)}{1-G(x)}\pi_0(dx)\Big)u(s)\,ds \qquad \text{a.e. } t \in (0,\infty). \tag{6.8}$$

Indeed, define

$$Y(t) \doteq \int_{[0,M)} \frac{G(x+t)-G(x)}{1-G(x)}\pi_0(dx), \qquad t \in [0,\infty),$$

and observe that the Lebesgue–Stieltjes measure dZ on $[0,\infty)$ is the convolution of the renewal measure U with the Lebesgue–Stieltjes measure dY. Moreover, by Fubini's theorem, for $t \in [0,\infty)$,

$$\begin{aligned} Y(t) &= \int_{[0,M)} \Big(\int_0^t \frac{g(x+s)}{1-G(x)}\,ds\Big)\pi_0(dx) \\ &= \int_0^t \Big(\int_{[0,M)} \frac{g(x+s)}{1-G(x)}\pi_0(dx)\Big)\,ds, \end{aligned}$$

which shows that dY is absolutely continuous with respect to Lebesgue measure and has density $y(\cdot) \doteq \int_{[0,M)} g(x+\cdot)/(1-G(x))\pi_0(dx)$. Since dU is also absolutely continuous with density u, it follows that (see, e.g., Problem 5 of Chapter 7 of [20]) dZ is also absolutely continuous with respect to Lebesgue measure, with density dZ/dt at t equal to the convolution of the functions y with u on $[0,t]$, for almost every $t \in [0,\infty)$. This proves (6.8).

Next, let $\tilde{\alpha}$ be the backward recurrence time process associated with a renewal process that has interrenewal distribution G and, for $x \in [0,M)$, let $\tilde{\mathbb{P}}_x$ be the law of $\tilde{\alpha}$ conditioned on $\tilde{\alpha}(0) = x$, let $\tilde{\mathbb{E}}_x$ denote the corresponding expectation and for any $\mu \in \mathcal{M}_{\le 1}[0,M)$, let $\tilde{\mathbb{E}}_\mu[\cdot] \doteq \int_{[0,M)} \tilde{\mathbb{E}}_x[\cdot]\mu(dx)$. We now show that

$$\langle f, \pi_t\rangle = \tilde{E}_{\pi_0}[f(\tilde{\alpha}_t)], \qquad f \in \mathcal{C}_b(\mathbb{R}_+), t \in [0,\infty). \tag{6.9}$$

Indeed, it is well known (see Proposition 1.5 and Example 2.1 of Chapter V in [1]) that $\{\tilde{\alpha}_s\}$ is a strong Markov process and that, for $f \in \mathcal{C}_b(\mathbb{R}_+)$, $\tilde{\mathbb{E}}_0[f(\tilde{\alpha}_t)]$ satisfies the renewal equation

$$\tilde{\mathbb{E}}_0[f(\tilde{\alpha}_t)] = \big(1-G(t)\big)f(t) + \int_0^t \tilde{\mathbb{E}}_0[f(\tilde{\alpha}_{t-s})]\,dG(s), \qquad t \in [0,\infty),$$

and hence (see Theorem 2.4 of Chapter V in [1]) admits the representation

$$\begin{aligned} \tilde{\mathbb{E}}_0[f(\tilde{\alpha}_t)] &= \int_0^t \big(1-G(t-s)\big)f(t-s)u(s)\,ds \\ &= (1-G)f \star u(t), \qquad t \in [0,\infty). \end{aligned}$$

Here, for any two functions f_1 and f_2, $f_1 \star f_2$ denotes the convolution of f_1 and f_2 on $[0,t]$. By standard renewal theory (see (3.1) on page 116 of [1]), if β_s is the forward recurrence time at $s \geq 0$, then $\mathbb{P}(\beta_0 > u | \tilde{\alpha}_0 = x) = (1 - G(x+u))/(1 - G(x))$. Moreover, $\tilde{\alpha}_t = \tilde{\alpha}_0 + t$ on the event $\{\beta_0 > t\}$ and, conditioned on $\beta_0 \in (s, s+ds)$, $\tilde{\alpha}_t$ under $\tilde{\mathbb{P}}_x$ has the same distribution as $\tilde{\alpha}_{t-s}$ under $\tilde{\mathbb{P}}_0$. Therefore, for $x \in [0, M)$ and $f \in \mathcal{C}_b(\mathbb{R}_+)$,

$$\begin{aligned} \tilde{\mathbb{E}}_x[f(\tilde{\alpha}_t)] &= \frac{1 - G(x+t)}{1 - G(x)} f(x+t) \\ &\quad + \int_0^t \tilde{\mathbb{E}}_0[f(\tilde{\alpha}_{t-s})] \frac{g(x+s)}{1 - G(x)}\, ds, \end{aligned} \tag{6.10}$$

and therefore

$$\begin{aligned} \tilde{\mathbb{E}}_{\pi_0}[f(\tilde{\alpha}_t)] &= \int_{[0,M)} \frac{1 - G(x+t)}{1 - G(x)} f(x+t) \pi_0(dx) \\ &\quad + \int_{[0,M)} \left(\int_0^t \tilde{\mathbb{E}}_0[f(\tilde{\alpha}_{t-s})] \frac{g(x+s)}{1 - G(x)}\, ds \right) \pi_0(dx) \\ &= \int_{[0,M)} \frac{1 - G(x+t)}{1 - G(x)} f(x+t) \pi_0(dx) \\ &\quad + \int_0^t \left(\int_{[0,M)} \frac{g(x+s)}{1 - G(x)} \pi_0(dx) \right) \big((1-G) f \star u\big)(t-s)\, ds. \end{aligned}$$

However, using (6.8), it is clear that the last term above equals

$$y \star \big((1-G) f \star u\big)(t) = (1-G) f \star (y \star u)(t) = (1-G) f \star \frac{dZ}{dt}(t),$$

which is equal to the last term in (6.6), and therefore (6.9) holds.

On the other hand, by the renewal theorem (see (4.5) of Theorem 4.3 of Chapter V in [1]), for every $x \in [0, M)$, $\lim_{t\to\infty} \tilde{\mathbb{E}}_x[f(\tilde{\alpha}_t)] = \langle f, \overline{\nu}_* \rangle$ and therefore $\lim_{t\to\infty} \langle f, \pi_t \rangle = \lim_{t\to\infty} \tilde{\mathbb{E}}_{\pi_0}[f(\tilde{\alpha}_t)] = \langle \mathbf{1}, \pi_0 \rangle \langle f, \overline{\nu}_* \rangle$, which is (6.7). $\square$

LEMMA 6.3. *If the service distribution has a finite second moment, then given any $\varepsilon > 0$, there exists $T_\varepsilon \in (0, \infty)$ such that*

$$\int_{[0,t]} \left(\int_{T+t-s}^{\infty} \big(1 - G(r)\big)\, dr \right) U(ds) \leq \varepsilon \qquad \text{for all } T \geq T_\varepsilon. \tag{6.11}$$

PROOF. Since the service distribution has a finite second moment, it follows that $\int_{[0,M)} r(1 - G(r)) < \infty$ and so there exists $\tilde{T}_\varepsilon < \infty$ (which we can always choose to be larger than M if $M < \infty$) such that

$$\int_T^{\infty} r\big(1 - G(r)\big)\, dr \leq \frac{\varepsilon}{2}, \qquad T \geq \tilde{T}_\varepsilon. \tag{6.12}$$

Now, for any $T < \infty$, we have

$$\begin{aligned}
&\int_{[0,t]} \left(\int_{T+t-s}^{\infty} \big(1 - G(r)\big)\, dr \right) U(ds) \\
&\quad = \int_{T}^{\infty} \left(\int_{[(T+t-r),t]} U(ds) \right) \big(1 - G(r)\big)\, dr \\
&\quad = \int_{T}^{T+t} \big(1 - G(r)\big)\big(U(t) - U(T+t-r)\big)\, dr \\
&\qquad + U(t) \int_{T+t}^{\infty} \big(1 - G(r)\big)\, dr.
\end{aligned}$$

We shall estimate the two terms on the last line above separately. Since G has a finite second moment, by Lorden's inequality (see Proposition V.6.2 of [1]) and the fact that $\int x\, dG(x)$, which is the mean time between renewals, equals 1, it follows that $U(t) - t$ is nonnegative and bounded by a constant, that we shall denote by B. Using this, along with the inequality $U(t) - U(t-r) \le U(r)$ for any $0 \le r \le t$, the first term can be bounded as follows: for $T \ge B$,

$$\begin{aligned}
&\int_{T}^{T+t} \big(1 - G(r)\big)\big(U(t) - U(T+t-r)\big)\, dr \\
&\quad = \int_{0}^{t} \big(1 - G(r+T)\big)\big(U(t) - U(t-r)\big)\, dr \\
&\quad \le \int_{0}^{t} \big(1 - G(r+T)\big) U(r)\, dr \\
&\quad \le \int_{0}^{t} \big(1 - G(r+T)\big)(r+B)\, dr \\
&\quad \le \int_{0}^{t} \big(1 - G(r+T)\big)(r+T)\, dr \\
&\quad \le \int_{T}^{\infty} r\big(1 - G(r)\big)\, dr.
\end{aligned}$$

As for the second term, we have for $T \ge B$,

$$\begin{aligned}
U(t) \int_{T+t}^{\infty} \big(1 - G(r)\big)\, dr &\le (t+B) \int_{T+t}^{\infty} \big(1 - G(r)\big)\, dr \\
&\le (t+T) \int_{T+t}^{\infty} \big(1 - G(r)\big)\, dr \\
&\le \int_{T+t}^{\infty} r\big(1 - G(r)\big)\, dr \\
&\le \int_{T}^{\infty} r\big(1 - G(r)\big)\, dr.
\end{aligned}$$

The last three displays, when combined with (6.12), show that the result holds with $T_\varepsilon = \max(B, \tilde{T}_\varepsilon)$. $\square$

PROOF OF THEOREM 3.9. The first statement of Theorem 3.9 follows from Proposition 6.1(2). For the second statement of the theorem, consider an arbitrary initial condition of the form $(\mathrm{id}, \overline{X}(0), \overline{\nu}_0) \in \mathcal{S}_0$, let $(\overline{X}, \overline{\nu})$ be the unique solution to the associated fluid equations (which exists by Theorem 3.7) and let $\overline{K}$ and $\overline{D}$ be the related processes defined in (3.8) and (3.9), respectively. Since $\overline{E} = \mathrm{id}$ is absolutely continuous, by Theorem 3.5 $\overline{K}$ is also absolutely continuous, with derivative $\overline{\kappa}$ that satisfies (3.12). Moreover, by Proposition 6.1(3) and the fact that $\langle \mathbf{1}, \overline{\nu}_t \rangle \le 1$ for all $t \in [0, \infty)$, shows that $\langle \mathbf{1}, \overline{\nu}_t \rangle \to 1$ as $t \to \infty$. We now consider the following two mutually exhaustive cases:

Case 1. There exists $T' \in (0, \infty)$ such that $\langle \mathbf{1}, \overline{\nu}_t \rangle < 1$ for all $t \ge T'$.

In this case, $\overline{X}(t) < 1$ for all $t \ge T'$. Therefore, by (3.12) it follows that $\overline{\kappa}(t) = \overline{\lambda} = 1$ for all $t \ge T'$. As a result, by Lemma 3.4 and (3.11), for any $r \ge 0$, we have

$$\langle f, \overline{\nu}_{T'+r} \rangle = \int_{[0,M)} f(x+r) \frac{1 - G(x+r)}{1 - G(x)} \overline{\nu}_{T'}(dx) + \int_0^r f(r-s)\big(1 - G(r-s)\big)\, ds.$$

Since f is uniformly bounded on $[0, \infty)$ and for every $x \in (0, \infty)$, $(1 - G(x + r))/(1 - G(x)) \to 0$ as $r \to \infty$, by the bounded convergence theorem, the first term converges to zero as $r \to \infty$. On the other hand, the second term trivially converges to $\int_0^\infty f(x)(1 - G(x))\, dx = \langle f, \overline{\nu}_* \rangle$ as $r \to \infty$. Since $\lim_{t\to\infty} \langle f, \overline{\nu}_t \rangle = \lim_{r\to\infty} \langle f, \overline{\nu}_{T'+r} \rangle$, this completes the proof of the theorem in this case.

Case 2. Given any $T' < \infty$, there exists $T > T'$ such that $\langle \mathbf{1}, \overline{\nu}_T \rangle = 1$.

Fix $\varepsilon > 0$. By Lemma 6.3, there exists $T_\varepsilon < \infty$ such that the estimate (6.11) holds for all $T \ge T_\varepsilon$. Choose $T \ge T_\varepsilon$ such that $\langle \mathbf{1}, \overline{\nu}_T \rangle = 1$ (which exists by the case assumption). Let $\pi_0 \doteq \overline{\nu}_T$, and let Z be defined as in Lemma 6.2. By the representation (4.6) for $\overline{K}$ given in Corollary 4.4 and the nonanticipative property of Lemma 3.4, it then follows that

$$\overline{K}^{[T]}(t) = \int_{[0,t]} \big(\langle \mathbf{1}, \overline{\nu}_{T+t-s} \rangle - \langle \mathbf{1}, \overline{\nu}_T \rangle\big) U(ds) + \int_{[0,t]} \left(\int_{[0,M)} \frac{G(x+t-s) - G(x)}{1 - G(x)} \overline{\nu}_T(dx) \right) U(ds), \tag{6.13}$$

where recall $\overline{K}^{[T]}(\cdot) = \overline{K}(T + \cdot)$. On the other hand, the comparison property (6.2) in Proposition 6.1(3) shows that for every $0 \le s \le t$,

$$0 \le \langle \mathbf{1}, \overline{\nu}_T \rangle - \langle \mathbf{1}, \overline{\nu}_{T+t-s} \rangle = 1 - \langle \mathbf{1}, \overline{\nu}_{T+t-s} \rangle \le \int_{T+t-s}^\infty \big(1 - G(r)\big)\, dr.$$

As a result, comparing the expression for Z given in (6.5) with that for $\overline{K}^{[T]}$ in (6.13), and using the last inequality and the estimate (6.11), we obtain for every $t \in [0, \infty)$,

$$\begin{aligned} 0 &\le \overline{K}^{[T]}(t) - Z(t) \\ &= \int_{[0,t]} (\langle \mathbf{1}, \overline{\nu}_T \rangle - \langle \mathbf{1}, \overline{\nu}_{T+t-s} \rangle) U(ds) \\ &\le \int_{[0,t]} \left(\int_{T+t-s}^{\infty} (1 - G(r))\, dr \right) U(ds) \le \varepsilon. \end{aligned} \tag{6.14}$$

Now, Lemma 3.4 and equation (3.11) show that for $f \in \mathcal{C}_b(\mathbb{R}_+)$, $\langle f, \overline{\nu}_{T+t} \rangle$ is equal to the right-hand side of the solution (4.3) to the age equation, but with $\overline{\nu}_0$ and Z replaced by $\overline{\nu}_T$ and $\overline{K}^{[T]}$, respectively, while (6.6) shows that $\langle f, \pi_t \rangle$ equals the right-hand side of (4.3), but with $\overline{\nu}_0$ replaced by $\overline{\nu}_T$ and Z as defined above. Therefore, by Lemma 4.5 and (6.14), we have for every $f \in \mathcal{C}_b[0, \infty)$,

$$\sup_{t \in [0,\infty)} |\langle f, \overline{\nu}_{T+t} \rangle - \langle f, \pi_t \rangle| \le C_f \sup_{t \in [0,\infty)} |\overline{K}^T(t) - Z(t)| \le C_f \varepsilon,$$

where $C_f = 2\|f\|_\infty + \|f'\|_\infty < \infty$. As an immediate consequence, we have

$$\limsup_{t \to \infty} |\langle f, \overline{\nu}_t \rangle - \langle f, \pi_t \rangle| \le C_f \varepsilon.$$

When combined with (6.7) of Lemma 6.2, this yields

$$\begin{aligned} \limsup_{t \to \infty} |\langle f, \overline{\nu}_t \rangle - \langle f, \overline{\nu}_* \rangle| &\le \limsup_{t \to \infty} |\langle f, \overline{\nu}_t \rangle - \langle f, \pi_t \rangle| \\ &\quad + \limsup_{t \to \infty} |\langle f, \pi_t \rangle - \langle f, \overline{\nu}_* \rangle| \\ &\le C_f \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this proves that $\overline{\nu}_t \xrightarrow{w} \overline{\nu}_*$ in Case 2 as well. This completes the proof of the theorem. □

Acknowledgments. We are very grateful to Luc Tartar for several useful discussions on the generalization of Theorem 4.1 to general h (see Section 4.3). We are also greatly indebted to the referees for a careful reading of the paper and, in particular, to one referee for a detailed and insightful report that greatly improved the exposition of the paper. In addition, we would like to thank Avi Mandelbaum for bringing this open problem to our attention and to thank Jim Dai, Weining Kang, Guodong Pang and Ward Whitt for their feedback on earlier versions of the paper. In particular, we are grateful to Weining Kang for a question that led to the discovery of an error in an earlier version of the proof of Theorem 3.8(2). Lastly, the second author would like to thank Chris Burdzy and Z.-Q. Chen for their hos-

pitality during her stay at the University of Washington in the Fall of 2006, when most of this work was completed.

REFERENCES

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

[2] 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

[3] DECREUSEFOND, L. and MOYAL, P. (2008). Fluid limit of a heavily loaded EDF queue with impatient customers. *Markov Process. Related Fields* **14** 131–158. MR2433299

[4] DUPUIS, P. and ELLIS, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York. MR1431744

[5] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. Wiley, New York. MR838085

[6] GROMOLL, H. C., PUHA, A. L. and WILLIAMS, R. J. (2002). The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.* **12** 797–859. MR1925442

[7] GROMOLL, H. C., ROBERT, P. and ZWART, B. (2008). Fluid limits for processor-sharing queues with impatience. *Math. Oper. Res.* **33** 375–402. MR2415999

[8] HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588. MR629195

[9] HARRISON, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley, New York. MR798279

[10] JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **288**. Springer, Berlin. MR959133

[11] JAKUBOWSKI, A. (1986). On the Skorokhod topology. *Ann. Inst. H. Poincaré Probab. Statist.* **22** 263–285. MR871083

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

[13] KANG, W. and RAMANAN, K. (2011). Asymptotic approximations for the stationary distributions of many-server queues. *Ann. Appl. Probab.* To appear.

[14] KASPI, H. and RAMANAN, K. (2010). SPDE limits of many server queues. Preprint.

[15] MANDELBAUM, A., MASSEY, W. A. and REIMAN, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Syst.* **30** 149–201. MR1663767

[16] PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. *Probability and Mathematical Statistics* **3**. Academic Press, New York. MR0226684

[17] PANG, G., TALREJA, R. and WHITT, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surv.* **4** 193–267. MR2368951

[18] RAMANAN, K. and REIMAN, M. I. (2003). Fluid and heavy traffic diffusion limits for a generalized processor sharing model. *Ann. Appl. Probab.* **13** 100–139. MR1951995

[19] REED, J. (2009). The $G/GI/N$ queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* **19** 2211–2269. MR2588244

[20] RUDIN, W. (1974). *Real and Complex Analysis*, 2nd ed. McGraw-Hill, New York. MR0344043

[21] SHILOV, G. E. (1968). *Generalized Functions and Partial Differential Equations*. Gordon and Breach, New York. MR0230129

[22] WHITT, W. (2006). Fluid models for multiserver queues with abandonments. *Oper. Res.* **54** 37–54. MR2201245

DEPARTMENT OF INDUSTRIAL ENGINEERING
AND MANAGEMENT
TECHNION—ISRAEL INSTITUTE OF TECHNOLOGY, HAIFA
ISRAEL
E-MAIL: iehaya@techunix.technion.ac.il

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