The Annals of Applied Probability
2010, Vol. 20, No. 6, 2204–2260
DOI: 10.1214/10-AAP683
© Institute of Mathematical Statistics, 2010

FLUID LIMITS OF MANY-SERVER QUEUES WITH RENEGING

BY WEINING KANG AND KAVITA RAMANAN[1]

University of Maryland, Baltimore County and Brown University

This work considers a many-server queueing system in which impatient customers with i.i.d., generally distributed service times and i.i.d., generally distributed patience times enter service in the order of arrival and abandon the queue if the time before possible entry into service exceeds the patience time. The dynamics of the system is represented in terms of a pair of measure-valued processes, one that keeps track of the waiting times of the customers in queue and the other that keeps track of the amounts of time each customer being served has been in service. Under mild assumptions, essentially only requiring that the service and reneging distributions have densities, as both the arrival rate and the number of servers go to infinity, a law of large numbers (or fluid) limit is established for this pair of processes. The limit is shown to be the unique solution of a coupled pair of deterministic integral equations that admits an explicit representation. In addition, a fluid limit for the virtual waiting time process is also established. This paper extends previous work by Kaspi and Ramanan, which analyzed the model in the absence of reneging. A strong motivation for understanding performance in the presence of reneging arises from models of call centers.

CONTENTS

Received December 2008; revised September 2009.

[1]Supported in part by the NSF Grants DMS-04-06191, CMMI-0728064 and CMMI-0928154. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

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

Key words and phrases. Many-server queues, GI/G/N queue, fluid limits, reneging, abandonment, strong law of large numbers, measure-valued processes, call centers.

1. Introduction.

1.1. *Background and motivation*. We consider a many-server queueing system in which customers with independent, identically distributed (henceforth, i.i.d.) service requirements chosen from a general distribution are processed in the order of arrival. In addition, a customer is assumed to abandon the queue if his/her time spent waiting in queue reaches his/her patience time. The patience times of customers are also assumed to be i.i.d. and drawn from a general distribution. When there are N servers and the cumulative customer arrival process is assumed to be a renewal process, this reduces to the so-called G/GI/N + GI model.

Over the last couple of decades, several applications have spurred the study of many-server models with abandonment [1, 3, 7]. Specifically, in applications to telephone contact centers and (more generally) customer contact centers, the effect of customers' impatience has been shown to have a substantial impact on the performance of the system [7]. For example, customer abandonment can stabilize a system that was formerly unstable. Under the assumption that the interarrival, service and abandonment time distributions are (possibly time-varying) exponential, process-level fluid and diffusion approximations were obtained by Mandelbaum, Massey and Reiman [16] for the total number in system in networks of multiserver queues with abandonments and retrials.

On the other hand, for the case of Poisson arrivals, exponential service times and general abandonment distributions (the M/M/N + GI queue), explicit formulae for the steady state distributions of the queue length and virtual waiting time were obtained by Baccelli and Hebuterne [1] (see Sections IV and V.2 therein), whereas several other steady state performance measures and their asymptotic approximations, in the limit as the arrival rates and servers go to infinity, were derived by Mandelbaum and Zeltyn [18]. In addition, approximations for performance measures suggested by these limit theorems were used by Garnett et al. [8] and Mandelbaum and Zeltyn [19] for the case of exponential and general abandonment distributions, respectively, to provide insight into the design of large call centers.

In all the previously mentioned works, the service times were assumed to be exponential. However, statistical analysis of real call centers has shown that both service times and abandon times are typically not exponentially distributed [4, 18], thus providing strong motivation for considering many-server systems with general service and abandonment distributions. A step toward incorporating more realistic general service distributions was taken in the insightful paper by Whitt [22], where a deterministic fluid approximation for a G/GI/N + GI queue with general service and abandonment distributions was proposed. However, the convergence of the discrete system starting empty to this fluid approximation was left as a conjecture (see Conjecture 2.1 in [22]). In this work, we rigorously identify the functional law of large numbers or mean-field limit, as the number of servers goes to infinity, of a many-server queueing system with general service and abandonment distributions starting from general initial conditions. In a recent work, Mandelbaum and

Momcilovic [17] have established diffusion approximations for the queue-length and virtual waiting time processes in a G/GI/N + GI queue.

With a view to providing a Markovian representation of the dynamics with a state space that is independent of the number of servers, we introduce a pair of measure-valued processes to describe the evolution of the system. One measure-valued process keeps track of the waiting times of customers in queue and the other keeps track of the amounts of time each customer present in the system has been in service. Under rather general assumptions (specified in Sections 2.1 and 3.1), we establish an asymptotic limit theorem for the scaled (divided by N) pair of measure-valued processes, as the number of servers N and the mean arrival rate into the system simultaneously go to infinity. In a recent independent study, Zhang [24] also considered the fluid limit for the same G/GI/N + GI system by using a measure-valued representation. His approach is based on tracking the "residual" service and patience times rather than tracking the "ages" in system and service as considered in this work. As in [14] and [15], an advantage of the particular measure-valued representation used here, in terms of ages in system and service, rather than residual service and residual patience times, is that it facilitates the application of martingale techniques, which streamlines the analysis and also allows for a more intuitive representation of the dynamics of the limiting process. In addition, the measure-valued approach also simultaneously allows for the characterization of asymptotic limits of several other functionals of interest. In order to illustrate this point, we also derive a limit theorem for the virtual waiting time of a customer, defined to be the time before entry to service of a (virtual) customer with infinite patience.

This work generalizes the framework of Kaspi and Ramanan [14], in which the corresponding model without abandonments was considered. The presence of two coupled measure-valued processes, rather than just one as in [14], makes the analysis here significantly more involved. In addition, an important step is the identification of an explicit expression for the cumulative reneging process. This paper also forms the basis of subsequent work in which we establish, under suitable conditions, the convergence of the stationary distributions of the fluid-scaled N-server systems to the invariant state of the fluid limit, as N tends to infinity [13].

It is worthwhile to mention that the models discussed above are relevant when the mean demand of customers is known (or can be accurately learned from an initial period of measurements), which is a realistic assumption in many applications. In other scenarios, it may be more natural to model the demand as being doubly stochastic. This approach was adopted by Harrison and Zeevi [9] (see also [2]), who proposed optimal staffing and design of multi-class call centers with several agent pools in the presence of abandonment under the assumption that the dominant variability arises from the randomness in the mean demand, rather than fluctuations around the mean demand.

1.2. *Outline of the paper.* The outline of the paper is as follows. We provide a more precise description of the model and the measure-valued representation of the state, and describe the dynamical equations governing the evolution of the system in Section 2 (the explicit construction of the state process is relegated to Appendix A and the strong Markov property of the state process is established in Appendix B). A key result here is Theorem 2.1, which provides a succinct characterization of the state dynamics. An analog of this characterization for continuous state processes leads to the fluid equations, which are introduced in Section 3.2 (see Definition 3.3). Next, the main results of the paper are summarized in Section 3.3. The first (Theorem 3.5) is a uniqueness result that states that (under the assumption that the service and abandonment distributions have densities and finite first moments) there exists at most one solution to the fluid equations. The proof of this result, which is considerably more involved than in the case without abandonment, is the subject of Section 4. The second and main result of the paper (Theorem 3.6) states that under mild additional assumptions (namely, Assumptions 3.1–3.3 introduced in Section 3.1), the scaled sequence of state processes converges weakly to the (unique) solution of the fluid equations, and provides a fairly explicit representation for the solution. The proof of this result consists of two main steps. First, in Section 6, the sequence of scaled state processes is shown to be tight and then, in Section 7, it is shown that every subsequential limit is a solution to the fluid equations. Both of these results make use of properties of a family of martingales that are established in Section 5. Finally, the last result (Theorem 3.8) formulates the asymptotic limit theorem for the virtual waiting time process, which is proved in Section 7.2. To start with, in Section 1.3, we first collect some basic notation and terminology used throughout the paper.

1.3. *Notation and terminology.* The following notation will be used throughout the paper. $\mathbb{Z}$ is the set of integers, $\mathbb{N}$ is the set of strictly positive integers, $\mathbb{R}$ is set of real numbers, $\mathbb{R}_+$ the set of nonnegative real numbers and $\mathbb{Z}_+$ is the set of nonnegative integers. 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].

1.3.1. *Function and measure spaces.* Given any metric space E, $\mathcal{C}_b(E)$ and $\mathcal{C}_c(E)$ are, respectively, the space of bounded, continuous functions and the space of continuous real-valued functions with compact support defined on E, while $\mathcal{C}^1(E)$ is the space of real-valued, once continuously differentiable functions on E, and $\mathcal{C}^1_c(E)$ is the subspace of functions in $\mathcal{C}^1(E)$ that have compact support. The subspace of functions in $\mathcal{C}^1(E)$ that, together with their first derivatives, are bounded, will be denoted by $\mathcal{C}^1_b(E)$. For $H \le \infty$, let $\mathcal{L}^1[0, H)$ and $\mathcal{L}^1_{\text{loc}}[0, H)$, respectively, represent the spaces of integrable and locally integrable functions on $[0, H)$, where a locally integrable function f on $[0, H)$ is a measurable function on $[0, H)$ that satisfies $\int_{[0,a]} f(x)\,dx < \infty$ for all $a < H$. The constant functions $f \equiv 1$ and $f \equiv 0$ will be represented by the symbols **1** and **0**, respectively. Given any càdlàg, real-valued function φ defined on $[0, \infty)$, we define $\|\varphi\|_T \doteq \sup_{s\in[0,T]} |\varphi(s)|$ for every $T < \infty$, and let $\|\varphi\|_\infty \doteq \sup_{s\in[0,\infty)} |\varphi(s)|$, which could possibly take the value ∞. In addition, the support of a function φ is denoted by $\mathrm{supp}(\varphi)$. Given a nondecreasing function f on $[0, \infty)$, f^{-1} denotes the inverse function of f in the sense that

$$f^{-1}(y) = \inf\{x \geq 0 : f(x) \geq y\}. \tag{1.1}$$

For each differentiable function f defined on $\mathbb{R}$, f' denotes the first derivative of f. For each function $f(t, x)$ defined on $\mathbb{R} \times \mathbb{R}^n$, f_t denotes the partial derivative of f with respect to t, and f_x denotes the partial derivative of f with respect to x.

The space of Radon measures on a metric space E, endowed with the Borel σ-algebra, is denoted by $\mathcal{M}(E)$, while $\mathcal{M}_F(E)$, $\mathcal{M}_1(E)$ and $\mathcal{M}_{\leq 1}(E)$ are, respectively, the subspaces of finite, probability and sub-probability measures in $\mathcal{M}(E)$. Also, given $B < \infty$, $\mathcal{M}_{\leq B}(E) \subset \mathcal{M}_F(E)$ denotes the space of measures μ in $\mathcal{M}_F(E)$ such that $|\mu(E)| \leq B$. Recall that a Radon measure is one that assigns finite measure to every relatively compact subset of $\mathbb{R}_+$. The space $\mathcal{M}(E)$ is equipped with the vague topology, that is, a sequence of measures $\{\mu_n\}$ in $\mathcal{M}(E)$ is said to converge to μ in the vague topology (denoted $\mu_n \stackrel{v}{\to} \mu$) if and only if for every $\varphi \in \mathcal{C}_c(E)$,

$$\int_E \varphi(x)\mu_n(dx) \to \int_E \varphi(x)\mu(dx) \qquad \text{as } n \to \infty. \tag{1.2}$$

By identifying a Radon measure $\mu \in \mathcal{M}(E)$ with the mapping on $\mathcal{C}_c(E)$ defined by

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

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

$$\left|\int_E \varphi(x)\mu(dx)\right| \leq L_\mathcal{K}\|\varphi\|_\infty \qquad \forall \varphi \in \mathcal{C}_c(E) \text{ with } \mathrm{supp}(\varphi) \subset \mathcal{K}.$$

On $\mathcal{M}_F(E)$, we will also consider the weak topology, that is, a sequence $\{\mu_n\}$ in $\mathcal{M}_F(E)$ is said to converge weakly to μ (denoted $\mu_n \stackrel{w}{\to} \mu$) if and only if (1.2) holds for every $\varphi \in \mathcal{C}_b(E)$. As is well known, $\mathcal{M}(E)$ and $\mathcal{M}_F(E)$, 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, by some abuse of notation, we will use **0** to denote the identically zero Radon measure on E. When E is an interval, say $[0, H)$, for notational conciseness, we will often write $\mathcal{M}[0, H)$ instead of $\mathcal{M}([0, H))$. For any finite measure μ on $[0, H)$, we define

$$F^\mu(x) \doteq \mu[0, x], \qquad x \in [0, H). \tag{1.3}$$

We say a measure μ is continuous at x if and only if $\mu(\{x\}) = 0$.

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

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

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

1.3.2. *Measure-valued stochastic processes*. Given a Polish space $\mathcal{H}$, we denote by $\mathcal{D}_{\mathcal{H}}[0, T]$ (resp., $\mathcal{D}_{\mathcal{H}}[0, \infty)$) the space of $\mathcal{H}$-valued, càdlàg functions on $[0, T]$ (resp., $[0, \infty)$), and we endow this space with the usual Skorokhod J_1-topology [20]. Then $\mathcal{D}_{\mathcal{H}}[0, T]$ and $\mathcal{D}_{\mathcal{H}}[0, \infty)$ are also Polish spaces (see [20]). In this work, we will be interested in $\mathcal{H}$-valued stochastic processes, where $\mathcal{H} = \mathcal{M}_F[0, H)$ for some $H \leq \infty$. 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 $\{X_n\}$ 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 $X_n \Rightarrow X$. Let $\mathcal{I}_{\mathbb{R}_+}[0, \infty)$ be the subset of nondecreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0, \infty)$ with $f(0) = 0$.

2. Description of model and state dynamics. In Section 2.1 we describe the basic model and the model primitives. In Section 2.2 we introduce the state descriptor and some auxiliary processes, and derive some equations that describe the dynamics of the state. Finally, in Section 2.3 (see Theorem 2.1), we provide a succinct characterization of the state dynamics. This characterization motivates the form of the fluid equations, which are introduced in Section 3.2.

2.1. *Model description and primitive data.* Consider a system with N servers, in which arriving customers are served in a nonidling, First-Come-First-Serve (FCFS) manner, 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, that there cannot simultaneously be a positive queue and an idle server, is satisfied. It is assumed that customers are impatient, and that a customer reneges from the queue as soon as the amount of time he/she has spent in queue reaches his/her patience time. Customers do not renege once they have entered service. The patience times of customers are given by an i.i.d. sequence, $\{r_i, i \in \mathbb{Z}\}$, with common cumulative distribution function G^r on $[0,\infty]$, while the service requirements of customers are given by another i.i.d. sequence, $\{v_i, i \in \mathbb{Z}\}$, with common cumulative distribution function G^s on $[0,\infty)$. For $i \in \mathbb{N}$, r_i and v_i represent, respectively, the patience time and the service requirement of the ith customer to enter the system after time zero, while $\{r_i, i \in -\mathbb{N} \cup \{0\}\}$ and $\{v_i, i \in -\mathbb{N} \cup \{0\}\}$ represent, respectively, the patience times and the service requirements of customers that arrived prior to time zero (if such customers exist), ordered according to their arrival times (prior to time zero). We assume that G^s has density g^s and G^r, restricted on $[0,\infty)$, has density g^r. This implies, in particular, that $G^r(0+) = G^s(0+) = 0$. Let

$$H^r \doteq \sup\{x \in [0,\infty) : g^r(x) > 0\},$$
$$H^s \doteq \sup\{x \in [0,\infty) : g^s(x) > 0\}$$

denote the right ends of the supports of g^r and g^s, respectively. The superscript (N) will be used to refer to quantities associated with the system with N servers.

Let $E^{(N)}$ denote the cumulative arrival process, with $E^{(N)}(t)$ representing the total number of customers that arrive into the system with N servers in the time interval $[0,t]$. Also, consider the càdlàg, real-valued process $\alpha_E^{(N)}$ defined by $\alpha_E^{(N)}(s) = s$ if $E^{(N)}(s) = 0$ and, if $E^{(N)}(s) > 0$, then

$$\alpha_E^{(N)}(s) \doteq s - \sup\{u < s : E^{(N)}(u) < E^{(N)}(s)\}, \tag{2.1}$$

which denotes the time elapsed since the last arrival. If $E^{(N)}$ is a renewal process, then $\alpha_E^{(N)}$ is simply the backward recurrence time process. Also, let $\mathcal{E}_0^{(N)}$ be an a.s. $\mathbb{Z}_+$-valued random variable that represents the number of customers that entered the system prior to time zero. This random variable does not play an important role in the analysis, but is used for bookkeeping purposes to keep track of the indices of customers.

The following mild assumptions on $E^{(N)}$ will be imposed throughout, without explicit mention:

- $E^{(N)}$ is a nondecreasing, pure jump process with $E^{(N)}(0) = 0$ and a.s., for $t \in [0, \infty)$, $E^{(N)}(t) < \infty$ and $E^{(N)}(t) - E^{(N)}(t-) \in \{0, 1\}$;
- the process $\alpha_E^{(N)}$ is Markovian with respect to its own natural filtration (this holds, e.g., when $E^{(N)}$ is a renewal process);
- the cumulative arrival process $E^{(N)}$, the sequence of service requirements $\{v_j, j \in \mathbb{Z}\}$ and the sequence of patience times $\{r_j, j \in \mathbb{Z}\}$ are independent.

The assumption on the jump size of $E^{(N)}$ is not crucial and is imposed mainly for convenience. On the other hand, the assumed independence of the service and patience times is a genuine restriction. It would be of interest to consider the case of correlated service and patience times.

2.2. *State descriptor and dynamical equations*. As mentioned in Section 1.1, our representation of the state of the system with N servers involves a pair of measure-valued processes, the "potential queue measure" process, $\eta^{(N)}$, which keeps track of the waiting times of customers in queue and the "age measure" process, $\nu^{(N)}$, which encodes the amounts of time that customers currently receiving service have been in service. In fact, the potential queue measure process keeps track not only of the waiting times of customers in queue, but also of the potential waiting times (equivalently, times since entry into system) of those customers who may have already entered service (and possibly departed the system), but for whom the time since entry into the system has not yet exceeded the patience time. In order to determine which subset of these customers is actually in queue, the process $X^{(N)}$, which represents the total number of customers in system with N servers (including those in service and those in queue), is also incorporated into the state descriptor. Thus the state of the system is represented by the vector of processes $(\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$, where $\alpha_E^{(N)}$ determines the cumulative arrival process via (2.1). The reason for introducing the process $\eta^{(N)}$ into the state (rather than working directly with a restricted measure that only encodes the waiting times of customers in queue) is that its dynamics is decoupled from the service dynamics. It is governed purely by the primitive data $E^{(N)}$ and G^r, and is thus more amenable to analysis (see Remark 2.2 for further elaboration of this point). Indeed, the queue measure process $\eta^{(N)}$ can also be viewed as describing the ages of customers in an infinite server queue that has cumulative arrivals $E^{(N)}$ and i.i.d. service requirements distributed according to G^r. Thus the dynamics of the process $\eta^{(N)}$ is also of independent interest.

Precise mathematical descriptions of $\eta^{(N)}$ and $\nu^{(N)}$ are given in Sections 2.2.1 and 2.2.2, respectively. Some auxiliary processes that are useful for describing the evolution of the state are introduced in Section 2.2.3. Finally, in Section 2.2.4, a filtration $\{\mathcal{F}_t^{(N)}\}$ corresponding to the system with N servers is introduced, and it is shown that the state processes and auxiliary processes are all adapted to this filtration. In fact, it is shown in Appendix B that the state process is a strong Markov process with respect to this filtration.

2.2.1. *Description of queue dynamics.* The potential waiting time process $w_j^{(N)}$ of customer j is (for every realization) defined to be the piecewise linear function on $[0,\infty)$ that is identically zero till the customer enters the system, then increases linearly, representing the amount of time elapsed since entering the system, and then remains constant (equal to the patience time) once the time elapsed exceeds the patience time. More precisely, for $j \in \mathbb{N}$, if $\zeta_j^{(N)}$ is the time at which the jth customer arrives into the system after time 0, then for $j \in \mathbb{N}$ $\zeta_j^{(N)} = (E^{(N)})^{-1}(j) \doteq \inf\{t > 0 : E^{(N)}(t) = j\}$ and

$$w_j^{(N)}(t) = \begin{cases} [t - \zeta_j^{(N)}] \vee 0, & \text{if } t - \zeta_j^{(N)} < r_j, \\ r_j, & \text{otherwise.} \end{cases} \tag{2.2}$$

For $j \in -\mathbb{N} \cup \{0\}$, $w_j^{(N)}$ represents the potential waiting time process of the jth customer who entered the system before time zero (if such a customer exists). Observe that the potential waiting time $w_j^{(N)}(t)$ of a customer at time t equals its actual waiting time (equivalently, time spent in queue) if and only if the customer has neither entered service nor reneged by time t. For $t \in [0,\infty)$, let $\eta_t^{(N)}$ be the nonnegative Borel measure on $[0, H^r)$ that has a unit mass at the potential waiting time of each customer that has entered the system by time t and whose potential waiting time has not yet reached its patience time. Recall that δ_x represents the Dirac mass at x. The potential queue measure $\eta_t^{(N)}$ can be written in the form

$$\begin{aligned} \eta_t^{(N)} &= \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \delta_{w_j^{(N)}(t)} \mathbb{1}_{\{w_j^{(N)}(t) < r_j\}} \\ &= \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \delta_{w_j^{(N)}(t)} \mathbb{1}_{\{dw_j^{(N)}/dt(t+) > 0\}}, \end{aligned} \tag{2.3}$$

where the last equality holds because at any time t, the potential waiting time process of any customer has a right derivative that is positive if and only if the customer has entered the system and the customer's potential waiting time has not yet reached its patience time.

For $t \in [0,\infty)$, let $Q^{(N)}(t)$ be the number of customers waiting in queue at time t. Due to the nonidling condition, the queue length process is then given by

$$Q^{(N)}(t) = [X^{(N)}(t) - N]^+. \tag{2.4}$$

Moreover, since the head-of-the-line customer is the customer in queue with the longest waiting time, the quantity

$$\chi^{(N)}(t) \doteq \inf\{x > 0 : \eta_t^{(N)}[0,x] \geq Q^{(N)}(t)\} = (F^{\eta_t^{(N)}})^{-1}(Q^{(N)}(t)) \tag{2.5}$$

represents the waiting time of the head-of-the-line customer in the queue at time t. Here, recall from (1.3) that $F^{\eta_t^{(N)}}$ is the c.d.f. of the measure $\eta_t^{(N)}$ and $(F^{\eta_t^{(N)}})^{-1}$ represents its inverse, as defined in (1.1). Since this is an FCFS system, any mass in $\eta_t^{(N)}$ that lies to the right of $\chi^{(N)}(t)$ represents a customer that has already entered service by time t. Therefore, the queue length process $Q^{(N)}$ admits the following alternative representation in terms of $\chi^{(N)}$ and $\eta^{(N)}$:

$$
\begin{aligned}
Q^{(N)}(t) &= \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \mathbb{1}_{\{w_j^{(N)}(t)\le\chi^{(N)}(t),\, w_j^{(N)}(t)<r_j\}} \\
&= \eta_t^{(N)}[0,\chi^{(N)}(t)].
\end{aligned}
\tag{2.6}
$$

2.2.2. *Description of service dynamics.* Analogous to the potential waiting process $w_j^{(N)}$, the age process $a_j^{(N)}$ associated with customer j is (for every realization) defined to be the piecewise linear function on $[0,\infty)$ that equals 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 completes service and departs the system. For $j=-\mathcal{E}_0^{(N)}+1,\dots,0$, let $a_j^{(N)}(0)$ represent the age of the jth customer in service at time 0 and for $j\in\mathbb{N}$, we set $a_j^{(N)}(0)=0$. Due to the First-Come-First-Serve (FCFS) nature of the system, customers in service at time t are those that did not renege, that have been in the system longer than the head-of-the-line customer at time t, but have not yet completed service and departed. Therefore, a.s., for $j=-\mathcal{E}_0^{(N)}+1,\dots,0,\dots,E^{(N)}(t)$, $t\ge 0$,

$$
\frac{da_j^{(N)}(t+)}{dt} = \begin{cases}
0, & \text{if } a_j^{(N)}(t)=0,\, w_j^{(N)}(t)=r_j, \\
 & \text{or } a_j^{(N)}(t)=0,\, w_j^{(N)}(t)\le\chi^{(N)}(t), \\
 & \text{or } a_j^{(N)}(t)=v_j, \\
1, & \text{if } a_j^{(N)}(t)=0,\, \chi^{(N)}(t)<w_j^{(N)}(t)<r_j, \\
 & \text{or } 0<a_j^{(N)}(t)<v_j.
\end{cases}
\tag{2.7}
$$

Note that the condition in the penultimate line of the right-hand side above represents the scenario in which a customer enters service precisely at time t, which causes $\chi^{(N)}$ to have a downward jump at time t since the condition that the arrival process increases only in unit jumps ensures that there is at most one customer with a given potential waiting time.

Now, for $t\in[0,\infty)$, let $\nu_t^{(N)}$ be the discrete nonnegative Borel measure on $[0,H^s)$ that has a unit mass at the age of each of the customers in service at time t.

Then, in a fashion analogous to (2.3), the age measure $\nu_t^{(N)}$ can be explicitly represented as

$$\nu_t^{(N)} = \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \delta_{a_j^{(N)}(t)} \mathbb{1}_{\{da_j^{(N)}/dt(t+)>0\}}. \tag{2.8}$$

2.2.3. *Auxiliary processes.* We now introduce certain auxiliary processes that will be useful for the study of the evolution of the system.

- The cumulative reneging process $R^{(N)}$, where $R^{(N)}(t)$ is the cumulative number of customers that have reneged from the system in the time interval $[0, t]$;
- the cumulative potential reneging process $S^{(N)}$, where $S^{(N)}(t)$ represents the cumulative number of customers whose potential waiting times have reached their patience times in the interval $[0, t]$;
- the cumulative departure process $D^{(N)}$, where $D^{(N)}(t)$ is the cumulative number of customers that have departed the system after completion of service in the interval $[0, t]$;
- the process $K^{(N)}$, where $K^{(N)}(t)$ represents the cumulative number of customers that have entered service in the interval $[0, t]$.

Now, a customer j completes service (and therefore departs the system) at time s if and only if, at time s, the left derivative of $a_j^{(N)}$ is positive and the right derivative of $a_j^{(N)}$ is zero. Therefore, we can write

$$D^{(N)}(t) = \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\{da_j^{(N)}/dt(s-)>0, da_j^{(N)}/dt(s+)=0\}}. \tag{2.9}$$

Note that the second sum in (2.9) is well defined since for each $t \geq 0$ and each j between $-\mathcal{E}_0^{(N)} + 1$ and $E^{(N)}(t)$, the piecewise linear structure of $a_j^{(N)}$ ensures that the indicator function in the sum is nonzero for at most one $s \in [0, t]$, that is, there exists at most one $s \in [0, t]$ such that the customer j completes service at time s. A similar logic shows that the cumulative potential reneging process $S^{(N)}$ admits the representation

$$S^{(N)}(t) = \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\{dw_j^{(N)}/dt(s-)>0, dw_j^{(N)}/dt(s+)=0\}}, \tag{2.10}$$

and the cumulative reneging process $R^{(N)}$ admits the representation

$$\begin{aligned} R^{(N)}(t) &\\ &= \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\{w_j^{(N)}(s)\leq\chi^{(N)}(s-), dw_j^{(N)}/dt(s-)>0, dw_j^{(N)}/dt(s+)=0\}}, \end{aligned} \tag{2.11}$$

where the additional restriction $w_j^{(N)}(s) \leq \chi^{(N)}(s-)$ is imposed so as to only count the reneging of customers actually in queue (and not the reneging of all customers in the potential queue, which is captured by $S^{(N)}$). Here, one considers the left limit $\chi^{(N)}(s-)$ of $\chi^{(N)}$ at time s to capture the situation in which $\chi^{(N)}$ jumps down at time s due to the head-of-the-line customer reneging from the queue or entering service.

Now, $\langle \mathbf{1}, \nu_t^{(N)} \rangle = \nu_t^{(N)}[0, \infty)$ represents the total number of customers in service at time t. Therefore, mass balances on the total number of customers in the system, the number of customers waiting in the "potential queue" and the number of customers in service show that

$$X^{(N)}(0) + E^{(N)} = X^{(N)} + D^{(N)} + R^{(N)}, \tag{2.12}$$

$$\langle \mathbf{1}, \eta_0^{(N)} \rangle + E^{(N)} = \langle \mathbf{1}, \eta^{(N)} \rangle + S^{(N)} \tag{2.13}$$

and

$$\langle \mathbf{1}, \nu_0^{(N)} \rangle + K^{(N)} = \langle \mathbf{1}, \nu^{(N)} \rangle + D^{(N)}. \tag{2.14}$$

In addition, it is also clear that

$$X^{(N)} = \langle \mathbf{1}, \nu^{(N)} \rangle + Q^{(N)}. \tag{2.15}$$

Combining (2.12), (2.14) and (2.15), we obtain the following mass balance for the number of customers in queue:

$$Q^{(N)}(0) + E^{(N)} = Q^{(N)} + R^{(N)} + K^{(N)}. \tag{2.16}$$

Furthermore, the nonidling condition takes the form

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

Indeed, note that this ensures that when $X^{(N)}(t) < N$, the number in the system is equal to the number in service, and so there is no queue, while if $X^{(N)}(t) > N$, there is a positive queue and $\langle \mathbf{1}, \nu_t^{(N)} \rangle = N$, indicating that there are no idle servers.

An advantage of the measure-valued state representation that we adopt is that it allows us to simultaneously study several other functionals of interest. As an example, we consider the so-called virtual waiting time process, which is important for applications. For each $t \geq 0$, the virtual waiting time $W^{(N)}(t)$ is defined to be the amount of time a (virtual) customer with infinite patience would have to wait before entering service if he were to arrive at time t. For a more precise definition of $W^{(N)}$, let $t \in [0, \infty)$ and for each $s \in [0, \infty)$, define

$$\begin{aligned} \mathcal{T}_t^{(N)}(s) \doteq \sum_{u \in [t, t+s]} \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} & \mathbb{1}_{\{dw_j^{(N)}/dt(u-)>0, dw_j^{(N)}/dt(u+)=0\}} \\ & \times \mathbb{1}_{\{w_j^{(N)}(u) \leq \chi^{(N)}(u-)\}}. \end{aligned} \tag{2.17}$$

Observe that $\mathcal{T}_t^{(N)}(s)$ equals the cumulative number of customers who arrived before time t and reneged from the queue (before entering service) in the time interval $[t, t+s]$. Once again, for each j there is at most one $u \in [t, t+s]$ for which both indicator functions in the summation are nonzero, and so the sum is well defined. The virtual waiting time $W^{(N)}(t)$ of a customer at time t is the least amount of time s that elapses after time t such that the cumulative departure from the system of customers that arrived prior to time t strictly exceeds the queue length at time t. Observing that this cumulative departure in the interval $[t, t+s]$ can be due to either departure from service or reneging of customers that arrived prior to time t, we can express the virtual waiting time as

$$W^{(N)}(t) \doteq \inf\{s \geq 0 : D^{(N)}(t+s) - D^{(N)}(t) + \mathcal{T}_t^{(N)}(s) > Q^{(N)}(t)\}. \tag{2.18}$$

Here, we have used the fact that for all s such that $D^{(N)}(t+s) - D^{(N)}(t) + \mathcal{T}_t^{(N)}(s) \leq Q^{(N)}(t)$, every customer that departed in the time interval $[t, t+s]$ must have arrived prior to time t.

2.2.4. *Filtration.* The total number of customers in service at time t is given by $\langle \mathbf{1}, \nu_t^{(N)} \rangle = \nu_t^{(N)}[0, H^s)$ and is bounded above by N. In addition, from (2.13) it follows that

$$\langle \mathbf{1}, \eta_t^{(N)} \rangle = \eta_t^{(N)}[0, H^r) \leq E^{(N)}(t) + \langle \mathbf{1}, \eta_0^{(N)} \rangle \leq E^{(N)}(t) + \mathcal{E}_0^{(N)},$$

which is a.s. finite by assumption. Therefore, for every $t \in [0, \infty)$, a.s., $\nu_t^{(N)} \in \mathcal{M}_F[0, H^s)$ and $\eta_t^{(N)} \in \mathcal{M}_F[0, H^r)$. Hence, the state descriptor $(\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$ takes values in $\mathbb{R}_+ \times \mathbb{Z}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$. For purely technical purposes we will find it convenient to also introduce the additional "station process" $s^{(N)} \doteq (s_j^{(N)}, j \in \mathbb{Z})$, 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 $s_j^{(N)}(t)$ is equal to the index $i \in \{1, \dots, N\}$ of the station at which customer j receives/received service and $s_j^{(N)}(t) \doteq 0$ otherwise. For $t \in [0, \infty)$, let $\tilde{\mathcal{F}}_t^{(N)}$ be the σ-algebra generated by

$$\{\mathcal{E}_0^{(N)}, X^{(N)}(0), \alpha_E^{(N)}(s), w_j^{(N)}(s), a_j^{(N)}(s), s_j^{(N)},$$
$$j \in \{-\mathcal{E}_0^{(N)} + 1, \dots, 0\} \cup \mathbb{N}, s \in [0, t]\},$$

and let $\{\mathcal{F}_t^{(N)}\}$ denote the associated right-continuous filtration, completed with respect to $\mathbb{P}$. In Appendix A, an explicit construction of the state descriptor and auxiliary processes is provided, which shows in particular that the state descriptor $(\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$ and auxiliary processes are càdlàg. Moreover, in Lemma A.1, it is proved that the state process $V^{(N)} \doteq (\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$ and the processes $E^{(N)}$, $Q^{(N)}$, $S^{(N)}$, $R^{(N)}$, $D^{(N)}$ and $K^{(N)}$ are all $\mathcal{F}_t^{(N)}$-adapted, and in Lemma B.1, it is shown that $(V^{(N)}, \mathcal{F}_t^{(N)})$ is a strong Markov process.

2.3. *A succinct characterization of the dynamics*. The main result of this section is Theorem 2.1, which provides equations that more succinctly characterize the dynamics of the state $(\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$ described in Section 2.2. First, we introduce some notation that is required to state the result.

For any measurable function φ on $[0, H^s) \times \mathbb{R}_+$, consider the process $D_\varphi^{(N)}$ that takes values in $\mathbb{R}$, and is given by

$$D_\varphi^{(N)}(t) \doteq \sum_{s\in[0,t]} \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \mathbb{1}_{\{da_j^{(N)}/dt(s-)>0, da_j^{(N)}/dt(s+)=0\}} \varphi\big(a_j^{(N)}(s), s\big) \tag{2.19}$$

for $t \in [0, \infty)$. It follows immediately from (2.19) and the right continuity of the filtration $\{\mathcal{F}_t^{(N)}\}$ that $D_\varphi^{(N)}$ is $\{\mathcal{F}_t^{(N)}\}$-adapted. Also, comparing (2.19) with (2.9), it is clear that when φ is the constant function $\mathbf{1}$, $D_\mathbf{1}^{(N)}$ is exactly the cumulative departure process $D^{(N)}$, that is,

$$D_\mathbf{1}^{(N)} = D^{(N)}. \tag{2.20}$$

In an exactly analogous fashion, for any measurable function ψ on $[0, H^r) \times \mathbb{R}_+$, consider the process $S_\psi^{(N)}$ that takes values in $\mathbb{R}$, and is given by

$$S_\psi^{(N)}(t) \doteq \sum_{s\in[0,t]} \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \mathbb{1}_{\{dw_j^{(N)}/dt(s-)>0, dw_j^{(N)}/dt(s+)=0\}} \psi\big(w_j^{(N)}(s), s\big). \tag{2.21}$$

It follows immediately from (2.21) and the right continuity of the filtration $\{\mathcal{F}_t^{(N)}\}$ that for $t \in [0, \infty)$, $S_\psi^{(N)}$ is $\{\mathcal{F}_t^{(N)}\}$-adapted. Moreover, $S_\mathbf{1}^{(N)}$ is clearly equal to the cumulative potential reneging process $S^{(N)}$, that is,

$$S_\mathbf{1}^{(N)} = S^{(N)}. \tag{2.22}$$

In addition, using (2.12), (2.15) and the nonnegativity of $Q^{(N)}$, $R^{(N)}$ and $\langle \mathbf{1}, \nu^{(N)} \rangle$, it follows that for any $t \in [0, \infty)$ and bounded, measurable φ,

$$\mathbb{E}\big[\big|D_\varphi^{(N)}(t)\big|\big] \le \|\varphi\|_\infty \mathbb{E}\big[X^{(N)}(0) + E^{(N)}(t)\big] < \infty \tag{2.23}$$

and likewise, for each $t \in [0, \infty)$ and bounded measurable ψ, (2.13) shows that

$$\mathbb{E}\big[\big|S_\psi^{(N)}(t)\big|\big] \le \|\psi\|_\infty \mathbb{E}\big[\langle \mathbf{1}, \eta_0^{(N)} \rangle + E^{(N)}(t)\big] < \infty. \tag{2.24}$$

Next, comparing (2.11) with (2.21), it is clear that the cumulative reneging process $R^{(N)}$ satisfies

$$R^{(N)}(t) = S_{\theta^{(N)}}^{(N)}(t), \qquad t \ge 0, \tag{2.25}$$

where $\theta^{(N)}$ is given by

$$\theta^{(N)}(x, s) = \mathbb{1}_{[x,\infty)}\big(\chi^{(N)}(s-)\big), \qquad x \in \mathbb{R}, s \ge 0. \tag{2.26}$$

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

THEOREM 2.1. *The processes* $(E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$ *a.s. satisfy the following coupled set of equations: for* $\varphi \in \mathcal{C}_c^1([0, H^s) \times \mathbb{R}_+)$ *and* $t \in [0, \infty)$,

$$
\begin{aligned}
\langle \varphi(\cdot, t), \nu_t^{(N)} \rangle = {} & \langle \varphi(\cdot, 0), \nu_0^{(N)} \rangle + \int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \nu_s^{(N)} \rangle \, ds \\
& - D_\varphi^{(N)}(t) + \int_{[0,t]} \varphi(0, s) \, dK^{(N)}(s),
\end{aligned} \tag{2.27}
$$

for $\psi \in \mathcal{C}_c^1([0, H^r) \times \mathbb{R}_+)$ *and* $t \in [0, \infty)$,

$$
\begin{aligned}
\langle \psi(\cdot, t), \eta_t^{(N)} \rangle = {} & \langle \psi(\cdot, 0), \eta_0^{(N)} \rangle + \int_0^t \langle \psi_x(\cdot, s) + \psi_s(\cdot, s), \eta_s^{(N)} \rangle \, ds \\
& - S_\psi^{(N)}(t) + \int_{[0,t]} \psi(0, s) \, dE^{(N)}(s),
\end{aligned} \tag{2.28}
$$

$$
X^{(N)}(t) = X^{(N)}(0) + E^{(N)}(t) - D_{\mathbf{1}}^{(N)}(t) - R^{(N)}(t), \tag{2.29}
$$

$$
N - \langle \mathbf{1}, \nu_t^{(N)} \rangle = [N - X^{(N)}(t)]^+, \tag{2.30}
$$

where $K^{(N)}$ *satisfies* (2.14), $R^{(N)}$ *satisfies* (2.25) *and* $D_\varphi^{(N)}$ *and* $S_\psi^{(N)}$ *are the processes defined in* (2.19) *and* (2.21), *respectively.*

REMARK 2.2. In the service dynamics, customer arrivals into service are governed by the process $K^{(N)}$, the random duration in service is determined by the distribution G^s and departures are represented by $D^{(N)}$. As captured by (2.27) and (2.28), the dynamics of the *potential queue* is exactly analogous, with the customer arrivals now governed by the process $E^{(N)}$, the random duration of stay in the potential queue determined by G^r and potential departures due to reneging represented by $S^{(N)}$. Moreover, *given* $K^{(N)}$, the dynamics of $\nu^{(N)}$ are exactly the same as in the case without abandonment, which was well studied in [14]. However, in the presence of reneging, there is a significantly more complicated coupling between $\nu^{(N)}$ and $K^{(N)}$, as captured by (2.29) and (2.30). In particular, this involves the cumulative reneging process $R^{(N)}$, which has no analogy with any quantity in the system without abandonments. Instead, as shown in the sequel [see Lemma 5.4, (5.17) and Proposition 7.2], we will exploit the representation (2.25) of $R^{(N)}$ in terms of the "known" quantity $S^{(N)}$ in order to characterize the limit of the scaled sequence of reneging processes.

PROOF OF THEOREM 2.1. The proof of (2.27) can be carried out in exactly the same way as the proof of (5.2) in Theorem 5.1 of [14], since the definition of $\nu^{(N)}$

in [14] is equivalent to the definition given in (2.8) here since $da_j^{(N)}(t+)/dt = 0$ for all $j > K^{(N)}(t)$ in [14]. For the reasons mentioned in Remark 2.2, the proof of (2.28) is also analogous except that the condition that each $\nu_t^{(N)}$ has total mass no greater than N is replaced by the argument below, which shows that each $\eta_t^{(N)}$ has finite mass. We know that for $k = 0, \dots, \lfloor nt \rfloor$,

$$\langle \mathbf{1}, \eta_{(k+1)/n}^{(N)} \rangle \leq E^{(N)}\left(\frac{k+1}{n}\right) + \langle \mathbf{1}, \eta_0^{(N)} \rangle \leq E^{(N)}(t+1) + \langle \mathbf{1}, \eta_0^{(N)} \rangle.$$

Thus, by taking the supremum over $k = 0, \dots, \lfloor nt \rfloor$, we have a.s.

$$\sup_{k=0,\dots,\lfloor nt \rfloor} \langle \mathbf{1}, \eta_{(k+1)/n}^{(N)} \rangle \leq E^{(N)}(t+1) + \mathcal{E}_0^{(N)} < \infty. \tag{2.31}$$

Equation (2.29) follows from (2.12), (2.20) and (2.25), while (2.30) is just the nonidling condition formulated in Section 2.2.3. □

3. Main results. In this section we summarize our main results. First, in Section 3.1, we introduce the fluid-scaled quantities and state our basic assumptions. Then, in Section 3.2, we introduce the so-called fluid equations, which provide a continuous analog of the characterization of the discrete model given in Theorem 2.1. In Section 3.3 we present our main theorems. In particular, we show that the fluid equations uniquely characterize the strong law of large numbers or "fluid" limit of the many-server system, as the number of servers goes to infinity.

3.1. *Fluid scaling and basic assumptions.* Consider the following scaled versions of the basic processes described in Section 2. For each $N \in \mathbb{N}$, the scaled version of the state descriptor $(\overline{\alpha}_E^{(N)}, \overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})$ is given by

$$\overline{\alpha}_E^{(N)}(t) \doteq \alpha_E^{(N)}(t), \qquad \overline{X}^{(N)}(t) \doteq \frac{X^{(N)}(t)}{N}, \tag{3.1}$$

$$\overline{\nu}_t^{(N)}(B) \doteq \frac{\nu_t^{(N)}(B)}{N}, \qquad \overline{\eta}_t^{(N)}(B) \doteq \frac{\eta_t^{(N)}(B)}{N}, \tag{3.2}$$

for $t \in [0, \infty)$ and any Borel subset B of $\mathbb{R}_+$. Analogously, define

$$\overline{I}^{(N)} \doteq \frac{I^{(N)}}{N} \qquad \text{for } I = E, D, K, Q, R, S, \mathcal{T}_t. \tag{3.3}$$

Recall that $\mathcal{I}_{\mathbb{R}_+}[0, \infty)$ is the subset of nondecreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0, \infty)$ with $f(0) = 0$, $H^s = \sup\{x \in [0, \infty) : g^s(x) > 0\}$ and $H^r = \sup\{x \in [0, \infty) : g^r(x) > 0\}$. Define

$$\mathcal{S}_0 \doteq \left\{ \begin{matrix} (e, x, \nu, \eta) \in \mathcal{I}_{\mathbb{R}_+}[0, \infty) \times \mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r) : \\ 1 - \langle \mathbf{1}, \nu \rangle = [1 - x]^+ \end{matrix} \right\}. \tag{3.4}$$

$\mathcal{S}_0$ serves as the space of possible input data for the fluid equations. Our goal is to identify the limit in distribution of the quantities $(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})$, as $N \to \infty$.

To this end, we impose some natural assumptions on the sequence of initial conditions $(\overline{E}^{(N)}, \overline{X}^{(N)}(0), \overline{\nu}_0^{(N)}, \overline{\eta}_0^{(N)})$.

ASSUMPTION 3.1 (Initial conditions). There exists an $\mathcal{S}_0$-valued random variable $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0)$ 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.;
3. $\overline{\nu}_0^{(N)} \stackrel{w}{\to} \overline{\nu}_0$ in $\mathcal{M}_F[0, H^s)$;
4. $\overline{\eta}_0^{(N)} \stackrel{w}{\to} \overline{\eta}_0$ in $\mathcal{M}_F[0, H^r)$, and $\mathbb{E}[\langle 1, \overline{\eta}_0^{(N)}\rangle] \to \mathbb{E}[\langle 1, \overline{\eta}_0\rangle] < \infty$.

REMARK 3.1. If the limits in (1) and (2) of Assumption 3.1 hold only in distribution rather than almost surely, then 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. Also, (1) and (4) of Assumption 3.1 and (2.30) imply that, for every $T \in [0,\infty)$,

$$\sup_{t\in[0,T]} \sup_N \mathbb{E}\big[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(t)\big] \le \mathbb{E}\big[1 + \langle 1, \overline{\eta}_0^{(N)}\rangle + \overline{E}^{(N)}(T)\big] < \infty. \tag{3.5}$$

The next assumption imposes some regularity conditions on $\overline{\eta}_0$ and $\overline{E}$.

ASSUMPTION 3.2. For each $t \ge 0$, if $\overline{\eta}_0(\{t\}) > 0$, then $\overline{\eta}_0(t, t+\varepsilon) > 0$ for every $\varepsilon > 0$ and if $\overline{E}(t) - \overline{E}(t-) > 0$, then $\overline{E}(t-) - \overline{E}(t-\varepsilon) > 0$ for every $\varepsilon > 0$.

REMARK 3.2. Assumption 3.2 is trivially satisfied if $\overline{\eta}_0$ and $\overline{E}$ are continuous, that is, $\overline{\eta}_0(\{t\}) = 0$ for all $t \ge 0$ and the function $\overline{E}$ is continuous.

In order to state our last assumption, define the hazard rate functions of G^r and G^s in the usual manner

$$h^r(x) \doteq \frac{g^r(x)}{1 - G^r(x)}, \qquad x \in [0, H^r), \tag{3.6}$$

$$h^s(x) \doteq \frac{g^s(x)}{1 - G^s(x)}, \qquad x \in [0, H^s). \tag{3.7}$$

It is easy to verify that h^r and h^s are locally integrable on $[0, H^r)$ and $[0, H^s)$, respectively.

ASSUMPTION 3.3. There exists $L^s < H^s$ such that h^s is either bounded or lower-semicontinuous on (L^s, H^s), and, likewise, there exists $L^r < H^r$ such that h^r is either bounded or lower-semicontinuous on (L^r, H^r).

3.2. *Fluid equations*. We now introduce the so-called fluid equations and provide some intuition as to why the limit of any sequence $(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})$ should be expected to be a solution to these equations. In Section 7, we provide a rigorous proof of this fact.

DEFINITION 3.3 (Fluid equations). The càdlàg function $(\overline{X}, \overline{\nu}, \overline{\eta})$ defined on $[0, \infty)$ and taking values in $\mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ is said to solve the *fluid equations* associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$ and the hazard rate functions h^r and h^s if and only if for every $t \in [0, \infty)$,

$$\int_0^t \langle h^r, \overline{\eta}_s \rangle \, ds < \infty, \qquad \int_0^t \langle h^s, \overline{\nu}_s \rangle \, ds < \infty, \tag{3.8}$$

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

$$\begin{aligned} \langle \varphi(\cdot, t), \overline{\nu}_t \rangle = \langle \varphi(\cdot, 0), \overline{\nu}_0 \rangle &+ \int_0^t \langle \varphi_s(\cdot, s), \overline{\nu}_s \rangle \, ds + \int_0^t \langle \varphi_x(\cdot, s), \overline{\nu}_s \rangle \, ds \\ &- \int_0^t \langle h^s(\cdot) \varphi(\cdot, s), \overline{\nu}_s \rangle \, ds + \int_0^t \varphi(0, s) \, d\overline{K}(s), \end{aligned} \tag{3.9}$$

where

$$\overline{K}(t) = \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_0 \rangle + \int_0^t \langle h^s, \overline{\nu}_s \rangle \, ds; \tag{3.10}$$

for every $\psi \in \mathcal{C}_c^1([0, H^r) \times \mathbb{R}_+)$

$$\begin{aligned} \langle \psi(\cdot, t), \overline{\eta}_t \rangle = \langle \psi(\cdot, 0), \overline{\eta}_0 \rangle &+ \int_0^t \langle \psi_s(\cdot, s), \overline{\eta}_s \rangle \, ds + \int_0^t \langle \psi_x(\cdot, s), \overline{\eta}_s \rangle \, ds \\ &- \int_0^t \langle h^r(\cdot) \psi(\cdot, s), \overline{\eta}_s \rangle \, ds + \int_0^t \psi(0, s) \, d\overline{E}(s); \end{aligned} \tag{3.11}$$

$$\overline{Q}(t) = \overline{X}(t) - \langle \mathbf{1}, \overline{\nu}_t \rangle; \tag{3.12}$$

$$\overline{Q}(t) \le \langle \mathbf{1}, \overline{\eta}_t \rangle; \tag{3.13}$$

$$\overline{R}(t) = \int_0^t \left(\int_0^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y)) \, dy \right) ds, \tag{3.14}$$

where we recall that $F^{\overline{\eta}_t}(x) = \overline{\eta}_t[0, x]$;

$$\overline{X}(t) = \overline{X}(0) + \overline{E}(t) - \int_0^t \langle h^s, \overline{\nu}_s \rangle \, ds - \overline{R}(t) \tag{3.15}$$

and

$$1 - \langle \mathbf{1}, \overline{\nu}_t \rangle = [1 - \overline{X}(t)]^+. \tag{3.16}$$

It immediately follows from (3.12) and (3.16) that for each $t \in [0, \infty)$,

$$\overline{Q}(t) = [\overline{X}(t) - 1]^+. \tag{3.17}$$

For future use, we also observe that (3.10), (3.12) and (3.15), when combined, show that for every $t \in [0, \infty)$,

$$\overline{Q}(0) + \overline{E}(t) = \overline{Q}(t) + \overline{K}(t) + \overline{R}(t). \tag{3.18}$$

We now provide an informal, intuitive explanation for the form of the fluid equations. Equations (3.10), (3.12) and (3.15) are simply mass conservation equations, that are fluid analogs of (2.14), (2.15) and (2.29), respectively, while (3.13) expresses a bound, whose analog clearly holds in the pre-limit, as can be seen from (2.6). The relation (3.16) is simply the fluid analog of the nonidling condition (2.30). Equations (3.9) and (3.11), which govern the evolution of the fluid age measure $\overline{\nu}$ and queue measure $\overline{\eta}$, respectively, are natural analogs of the pre-limit equations (2.27) and (2.28), respectively. It is worthwhile to comment further on the fourth terms on the right-hand sides of (3.9) and (3.11), which characterize the fluid departure rate and potential reneging rate, respectively, as integrals of the corresponding hazard rate with respect to the age and queue measures. Note that $\overline{\nu}_s(dx)$ represents the amount of mass (limiting fraction of customers) whose age lies in the range $[x, x + dx)$ at time s, and $h^s(x)$ represents the fraction of mass with age x (i.e., with service time no less than x) that would depart from the system while having age in $[x, x + dx)$. Hence, it is natural to expect $\langle h^s, \overline{\nu}_s\rangle$ to represent the departure rate of mass from the fluid system at time s. This was rigorously proved in the case without abandonments in [14] (see Proposition 5.17 therein). By exploiting the exact analogy between $(\overline{\nu}, \overline{K}, \overline{D})$ and $(\overline{\eta}, \overline{E}, \overline{S})$ (see Remark 2.2), it is clear that the potential reneging rate at time s can be similarly represented as $\langle h^r, \overline{\eta}_s\rangle$. Thus the fluid potential reneging process $\overline{S}$, defined by

$$\overline{S}(t) \doteq \int_0^t \langle h^r, \overline{\eta}_s\rangle\, ds, \qquad t \in [0, \infty), \tag{3.19}$$

represents the cumulative amount of potential reneging from the fluid system in the interval $[0, t]$. Due to the FCFS nature of the system, the fluid queue at time s contains all the mass in $\overline{\eta}$ that is to the left of $(F^{\overline{\eta}_s})^{-1}(\overline{Q}(s))$, where recall $F^{\overline{\eta}_s}$ is the c.d.f. of $\overline{\eta}_s$. Moreover, roughly speaking, given any $y \in [0, \overline{Q}(s)]$, there is a mass of dy customers in the queue whose waiting time at s is $(F^{\overline{\eta}_s})^{-1}(y)$ and the mean reneging rate of customers with this waiting time is $h^r((F^{\overline{\eta}_s})^{-1}(y))$. Thus the total actual reneging that has occurred in the interval $[0, t]$, is represented by the integral, as specified in (3.14).

We close the section with a simple result on the action of time-shifts on solutions to the fluid equations. For this, we need the following notation: for any $t \in [0, \infty)$,

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

LEMMA 3.4. *Suppose the càdlàg function* $(\overline{X}, \overline{\nu}, \overline{\eta})$ *defined on* $[0,\infty)$ *and taking values in* $\mathbb{R}_+ \times \mathcal{M}_F[0,H^s) \times \mathcal{M}_F[0,H^r)$ *solves the fluid equations associated with* $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$, *then* $(\overline{X}^{[t]}, \overline{\nu}^{[t]}, \overline{\eta}^{[t]})$ *solves the fluid equations associated with* $(\overline{E}^{[t]}, \overline{X}(t), \overline{\nu}_t, \overline{\eta}_t) \in \mathcal{S}_0$, *where* $\overline{K}^{[t]}, \overline{R}^{[t]}, \overline{Q}^{[t]}$ *are the corresponding processes that satisfy* (3.10), (3.14), (3.12) *with* $\overline{\nu}^{[t]}, \overline{\eta}^{[t]}$ *and* $\overline{X}^{[t]}$ *in place of* $\overline{\nu}, \overline{\eta}$ *and* $\overline{X}$.

The proof of the lemma just involves a rewriting of the fluid equations, and is thus omitted.

3.3. *Summary of main results*. Our first result establishes uniqueness of solutions to the fluid equations.

THEOREM 3.5. *Given any* $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$, *there exists at most one solution* $(\overline{X}, \overline{\nu}, \overline{\eta})$ *to the associated fluid equations* (3.8)–(3.16). *Moreover, if* $\overline{\nu}$ *and* $\overline{\eta}$ *satisfy* (3.8), *then* $(\overline{X}, \overline{\nu}, \overline{\eta})$ *is a solution to the fluid equations if and only if for every* $f \in \mathcal{C}_b(\mathbb{R}_+)$,

$$\begin{aligned}\int_{[0,H^r)} f(x)\overline{\eta}_t(dx) &= \int_{[0,H^r)} f(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\overline{\eta}_0(dx) \\ &\quad + \int_{[0,t]} f(t-s)\bigl(1-G^r(t-s)\bigr)\,d\overline{E}(s),\end{aligned} \tag{3.20}$$

$$\begin{aligned}\int_{[0,H^s)} f(x)\overline{\nu}_t(dx) &= \int_{[0,H^s)} f(x+t)\frac{1-G^s(x+t)}{1-G^s(x)}\overline{\nu}_0(dx) \\ &\quad + \int_{[0,t]} f(t-s)\bigl(1-G^s(t-s)\bigr)\,d\overline{K}(s),\end{aligned} \tag{3.21}$$

where

$$\begin{aligned}\overline{K}(t) &= [\overline{X}(0)-1]^+ - [\overline{X}(t)-1]^+ + \overline{E}(t) \\ &\quad - \int_0^t \left(\int_0^{[\overline{X}(s)-1]^+} h^r((F^{\overline{\eta}_s})^{-1}(y))\,dy\right)ds\end{aligned} \tag{3.22}$$

and for all $t \in [0,\infty)$, $\overline{X}$ *satisfies* $[\overline{X}(t)-1]^+ \le \langle \mathbf{1}, \overline{\eta}_t\rangle$, *the nonidling condition* (3.16) *and*

$$\begin{aligned}\overline{X}(t) &= \overline{X}(0) + \overline{E}(t) - \int_0^t \langle h^s, \overline{\nu}_s\rangle\,ds \\ &\quad - \int_0^t \left(\int_0^{[\overline{X}(s)-1]^+} h^r((F^{\overline{\eta}_s})^{-1}(y))\,dy\right)ds.\end{aligned} \tag{3.23}$$

Moreover, $\overline{K}$ also satisfies

$$\begin{aligned}\overline{K}(t) &= \langle \mathbf{1}, \overline{\nu}_{t-s}\rangle - \langle \mathbf{1}, \overline{\nu}_0\rangle + \int_{[0,H^s)} \frac{G^s(x+t-s) - G^s(x)}{1-G^s(x)} \overline{\nu}_0(dx) \\ &\quad + \int_0^t \biggl(\langle \mathbf{1}, \overline{\nu}_{t-s}\rangle - \langle \mathbf{1}, \overline{\nu}_0\rangle \\ &\qquad + \int_{[0,H^s)} \frac{G^s(x+t-s) - G^s(x)}{1-G^s(x)} \overline{\nu}_0(dx) \biggr) u^s(s)\, ds,\end{aligned} \tag{3.24}$$

where u^s is the density of the renewal function U^s associated with G^s (u^s exists since G^s is assumed to have a density).

Next, we state the main result of the paper, which shows that, under fairly general conditions, a solution to the fluid equations exists and is the functional law of large numbers limit, as $N \to \infty$, of the N-server system with abandonment.

THEOREM 3.6. *Suppose that Assumptions 3.1–3.3 hold, and let $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$ be the limiting initial condition. Then there exists a unique solution $(\overline{X}, \overline{\nu}, \overline{\eta})$ to the associated fluid equations, and the sequence $(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})$ converges weakly, as $N \to \infty$, to $(\overline{X}, \overline{\nu}, \overline{\eta})$.*

Theorem 3.6 follows from Theorem 6.1, which establishes tightness of the sequence $\{\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)}\}$, Theorem 7.1, which shows that any subsequential limit of the sequence $\{\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)}\}$ satisfies the fluid equations, and the uniqueness of solutions to the fluid equations stated in Theorem 3.5.

COROLLARY 3.7. *Suppose that Assumptions 3.1–3.3 hold. Given any $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$, let $(\overline{X}, \overline{\nu}, \overline{\eta})$ be the unique solution to the associated fluid equations (3.8)–(3.16) specified in Theorem 3.5. If the function $\overline{E}$ is absolutely continuous and $\overline{\nu}_0$ and $\overline{\eta}_0$ are absolutely continuous measures, then the function $\overline{X}$ is also absolutely continuous and for every $t \in [0, \infty)$, the measures $\overline{\nu}_t$ and $\overline{\eta}_t$ are also absolutely continuous.*

PROOF. Since $\overline{E}$ is absolutely continuous, (3.23) allows us to deduce that $\overline{X}$ is absolutely continuous. In turn, (3.22) shows that $\overline{K}$ is also absolutely continuous. Then the argument used in proving Lemma 5.18 of [14] can be adapted, together with (3.20) and (3.21), to show that $\overline{\nu}_t$ and $\overline{\eta}_t$ are absolutely continuous for every $t \in [0, \infty)$. This proves the corollary. □

We now state the fluid limit result for the virtual waiting time process $W^{(N)}$. This result is of particular interest in the context of call centers. Note that in the fluid system, for any $u > t$ the total mass of customers in queue at time u that arrived before time t equals $\overline{Q}(u) - \overline{\eta}_u[0, u-t]$, and the ages of these (fluid) customers lie in the interval $(u-t, \overline{\chi}(u)]$, where

$$\overline{\chi}(u) \doteq (F^{\overline{\eta}_u})^{-1}(\overline{Q}(u)). \tag{3.25}$$

Observe that this definition is analogous to the definition of $\chi^{(N)}$ given in (2.5). Therefore, by the same logic that was used to justify the expression (3.14) for $\overline{R}$ in Definition 3.3, it is natural to conjecture that, for each $t \in [0, \infty)$, the fluid limit of the sequence $\{\overline{\mathcal{T}}_t^{(N)}\}$ equals $\overline{\mathcal{T}}_t$, where for $s \in [0, \infty)$,

$$\begin{aligned} \overline{\mathcal{T}}_t(s) &\doteq \int_t^{t+s} \left(\int_{\overline{\eta}_u[0,u-t]}^{\overline{Q}(u)} h^r((F^{\overline{\eta}_u})^{-1}(y))\,dy \right) du \\ &= \int_0^s \left(\int_{\overline{\eta}_{t+u}[0,u]}^{\overline{Q}(t+u)} h^r((F^{\overline{\eta}_{t+u}})^{-1}(y))\,dy \right) du. \end{aligned} \tag{3.26}$$

Also, define

$$\overline{W}(t) \doteq \inf\left\{ s \geq 0 : \int_t^{t+s} \langle h^s, \overline{\nu}_u \rangle\,du + \overline{\mathcal{T}}_t(s) \geq \overline{Q}(t) \right\}. \tag{3.27}$$

We will say a function $f \in \mathcal{D}[0, \infty)$ is uniformly strictly increasing if it is absolutely continuous and there exists $a > 0$ such that the derivative of f is bigger than and equal to a for a.e. $t \in [0, \infty)$. Note that for any such function, $f^{-1}(f(t)) = t$ and f^{-1} is continuous and strictly increasing on $[0, \infty)$. We now characterize the fluid limit of the (scaled) virtual waiting time in the system.

THEOREM 3.8. *Suppose that the conditions of Theorem 3.6 hold and that the function $\int_0^{\cdot} \langle h^s, \overline{\nu}_u \rangle\,du$ is uniformly strictly increasing. For each $t \geq 0$, if $\overline{Q}$ is continuous at t, then $\overline{\mathcal{T}}_t^{(N)} \Rightarrow \overline{\mathcal{T}}_t$ and $W^{(N)}(t) \Rightarrow \overline{W}(t)$, as $N \to \infty$.*

4. Uniqueness of solutions to the fluid equations. In Section 4.1, we show that if $(\overline{X}, \overline{\nu}, \overline{\eta})$ solve the fluid equations associated with a given initial condition $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$, then $\overline{\nu}$ (resp., $\overline{\eta}$) can be written explicitly in terms of the auxiliary fluid process $\overline{K}$ (resp., cumulative arrival process $\overline{E}$). In Section 4.2, these representations are used, along with the nonidling condition and the remaining fluid equations, to show that there is at most one solution to the fluid equations for a given initial condition.

4.1. *Integral equations for $(\overline{\nu}, \overline{K})$ and $(\overline{\eta}, \overline{E})$.* We begin by recalling Theorem 4.1 and Remark 4.3 of [14], which we state here as Proposition 4.1. This proposition identifies an implicit relation that must be satisfied by the processes $(\overline{\nu}, \overline{K})$ and $(\overline{\eta}, \overline{E})$ that solve (3.9) and (3.11), respectively.

PROPOSITION 4.1 [14]. *Let G be the cumulative distribution function of a probability distribution with density g and hazard rate function $h = g/(1-G)$, let $H \doteq \sup\{x \in [0,\infty) : g(x) > 0\}$. Suppose $\overline{\pi} \in \mathcal{D}_{\mathcal{M}_F[0,H)}[0,\infty)$ has the property that for every $m \in [0,H)$ and $T \in [0,\infty)$, there exists $C(m,T) < \infty$ such that*

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

for every $\varphi \in \mathcal{C}_c(\mathbb{R}^2)$ with $\operatorname{supp}(\varphi) \subset [0,m] \times [0,T]$. Then given any $\overline{\pi}_0 \in \mathcal{M}_F[0,H)$ and $\overline{Z} \in \mathcal{I}_{\mathbb{R}_+}[0,\infty)$, $\overline{\pi}$ satisfies the integral equation

$$\begin{aligned} \langle \varphi(\cdot,t), \overline{\pi}_t \rangle &= \langle \varphi(\cdot,0), \overline{\pi}_0 \rangle + \int_0^t \langle \varphi_s(\cdot,s), \overline{\pi}_s \rangle \, ds + \int_0^t \langle \varphi_x(\cdot,s), \overline{\pi}_s \rangle \, ds \\ &\quad - \int_0^t \langle \varphi(\cdot,s)h(\cdot), \overline{\pi}_s \rangle \, ds + \int_{[0,t]} \varphi(0,s) \, d\overline{Z}(s) \end{aligned} \tag{4.2}$$

for every $\varphi \in \mathcal{C}_c((-\infty,H) \times \mathbb{R})$ and $t \in [0,\infty)$, if and only if $\overline{\pi}$ satisfies

$$\begin{aligned} \int_{[0,M)} f(x)\overline{\pi}_t(dx) &= \int_{[0,M)} f(x+t) \frac{1-G(x+t)}{1-G(x)} \overline{\pi}_0(dx) \\ &\quad + \int_{[0,t]} f(t-s)\big(1-G(t-s)\big) \, d\overline{Z}(s), \end{aligned} \tag{4.3}$$

for every $f \in \mathcal{C}_b(\mathbb{R}_+)$ and $t \in (0,\infty)$. Moreover, for every $f \in \mathcal{C}_b^1(\mathbb{R}_+)$ and $t \in (0,\infty)$,

$$\begin{aligned} &\int_0^t f(t-s)\big(1-G(t-s)\big) \, d\overline{Z}(s) \\ &\quad = f(0)\overline{Z}(t) + \int_{[0,t]} f'(t-s)\big(1-G(t-s)\big)\overline{Z}(s) \, ds \\ &\qquad - \int_{[0,t]} f(t-s)g(t-s)\overline{Z}(s) \, ds. \end{aligned} \tag{4.4}$$

Fluid equations (3.8)–(3.11) show that (4.1) and (4.2) are satisfied with $(h, \overline{\pi}, \overline{Z})$ replaced by $(h^s, \overline{\nu}, \overline{K})$ and $(h^r, \overline{\eta}, \overline{E})$, respectively. Therefore, the next result follows from Proposition 4.1.

COROLLARY 4.2. *Processes $(\overline{\eta}, \overline{E})$ and $(\overline{\nu}, \overline{K})$ satisfy (3.20) and (3.21) for every bounded Borel measurable function f and $t \in [0,\infty)$. Moreover, $\overline{K}$ satisfies the renewal equation*

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

for each $t \geq 0$ *and admits the representation*

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

where dU^s *is the renewal measure associated with the distribution* G^s.

REMARK 4.3. Strictly speaking, in [14] the cumulative distribution function G was assumed to be absolutely continuous and supported on $[0, \infty)$. However, the proofs given there only use the local integrability of the hazard rate function h on $[0, H)$ and so continue to hold for G^r here, which may possibly have a positive mass at ∞. In fact, in the case that G^r has a positive mass at ∞ the hazard rate function h^r is globally integrable on $[0, H^r)$.

4.2. *Uniqueness of solutions.* Let $(\overline{X}, \overline{\nu}, \overline{\eta})$ be a solution to the fluid equations associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0)$. Recall the definitions of $\overline{Q}$ and $\overline{R}$ that are given in (3.12) and (3.14). As an immediate consequence of (3.14), we have the following elementary property.

LEMMA 4.4. *For any* $0 \leq a \leq b < \infty$, *if* $\overline{Q}(t) = 0$ *for all* $t \in [a, b]$, *then* $\overline{R}(b) - \overline{R}(a) = 0$.

Next, we establish the intuitive result that the process $\overline{K}$ that represents the cumulative entry of "fluid" into service is nondecreasing.

LEMMA 4.5. *The function* $\overline{K}$ *is nondecreasing.*

PROOF. Fix $t \in [0, \infty)$ and $0 \leq s < t$. If $\overline{X}(t) \geq 1$, then $\langle \mathbf{1}, \overline{\nu}_t\rangle = 1 \geq \langle \mathbf{1}, \overline{\nu}_s\rangle$ by (3.16). Hence, by (3.10), it follows that

$$\overline{K}(t) - \overline{K}(s) = \langle \mathbf{1}, \overline{\nu}_t\rangle - \langle \mathbf{1}, \overline{\nu}_s\rangle + \int_s^t \langle h^s, \overline{\nu}_l\rangle\, dl \geq 0. \tag{4.6}$$

If $\overline{X}(t) < 1$, we consider two cases.

Case 1. $\overline{X}(v) < 1$ for all $v \in (s, t]$. In this case, by (3.12) and (3.16), $\overline{Q}(v) = 0$ for all $v \in (s, t]$. Hence, by Lemma 4.4 and the right continuity of $\overline{R}$, $\overline{R}(t) - \overline{R}(s) = 0$. By (3.18), it then follows that

$$\begin{aligned}\overline{K}(t) - \overline{K}(s) &= \overline{K}(t) - \overline{K}(s) + \overline{R}(t) - \overline{R}(s) + \overline{Q}(t) - \overline{Q}(s) \\ &= \overline{E}(t) - \overline{E}(s) \\ &\geq 0.\end{aligned}$$

Case 2. There exists $v \in (s,t]$ such that $\overline{X}(v) \geq 1$. Define $l \doteq \sup\{v \leq t : \overline{X}(v) \geq 1\}$. Then, clearly $l \in (s,t]$ and $\overline{X}(l-) \geq 1$. Now, (3.14) implies that $\overline{R}$ is continuous and hence, by (3.15), $\overline{X}(v) - \overline{X}(v-) \geq 0$ for every $v \in (0,\infty)$. Therefore, $\overline{X}(l) \geq 1 = \langle \mathbf{1}, \overline{\nu}_l \rangle$, with the latter equality being a consequence of the nonidling condition (3.16). Due to the case assumption $\overline{X}(t) < 1$, we must have $l < t$. Then (4.6), with t replaced by l, shows that $\overline{K}(l) - \overline{K}(s) \geq 0$. On the other hand, since $\overline{X}(v) < 1$ for all $v \in (l,t]$, the argument in case 1 above shows that $\overline{K}(t) - \overline{K}(l) \geq 0$. Thus, in this case too, we have $\overline{K}(t) - \overline{K}(s) \geq 0$. $\square$

We now state the main result of this section.

THEOREM 4.6. *For $i = 1,2$, let $(\overline{X}^i, \overline{\nu}^i, \overline{\eta}^i)$ be a solution to the fluid equations associated with $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$. Then $\overline{X}^1 = \overline{X}^2$, $\overline{\nu}^1 = \overline{\nu}^2$ and $\overline{\eta}^1 = \overline{\eta}^2$.*

PROOF. For each $i = 1,2$, let $\overline{Q}^i, \overline{K}^i, \overline{D}^i, \overline{R}^i$ be the processes associated with the solution $(\overline{X}^i, \overline{\nu}^i, \overline{\eta}^i)$ to the fluid equations for $(\overline{E}, \overline{X}(0), \overline{\nu}_0, \overline{\eta}_0) \in \mathcal{S}_0$. It follows directly from Corollary 4.2 that $\overline{\eta}^1 = \overline{\eta}^2$. Let $\triangle A$ denote $A^2 - A^1$ for $A = \overline{Q}, \overline{K}, \overline{D}, \overline{R}$ and $\overline{\nu}$. For each $t \geq 0$, let $\triangle \overline{\nu}_t$ be the measure that satisfies $\triangle \overline{\nu}_t(\Xi) = \overline{\nu}_t^2(\Xi) - \overline{\nu}_t^1(\Xi)$ for every measurable set $\Xi \subset [0,\infty)$. Choose $\delta > 0$ and define

$$\tau = \tau_\delta \doteq \inf\{t \geq 0 : \triangle \overline{K}(t) \vee \triangle \overline{K}(t-) \geq \delta\}.$$

We shall argue by contradiction to show that $\tau = \infty$. Suppose that $\tau < \infty$.

We first claim that for each $t \in [0,\tau]$,

$$\triangle \overline{K}(t) < \delta \qquad \text{if } \langle \mathbf{1}, \overline{\nu}_t^1 \rangle = 1. \tag{4.7}$$

To see why this is true, suppose that $\langle \mathbf{1}, \overline{\nu}_t^1 \rangle = 1$ for some $t \in [0,\tau]$. Since $\langle \mathbf{1}, \overline{\nu}_t^2 \rangle \leq 1$, we have $\langle \mathbf{1}, \triangle \overline{\nu}_t \rangle \leq 0$. When combined with (4.5) and the identity $\triangle \overline{\nu}_0 = 0$, this shows that

$$\triangle \overline{K}(t) = \langle \mathbf{1}, \triangle \overline{\nu}_t \rangle + \int_0^t g^s(t-s) \triangle \overline{K}(s)\,ds \leq \int_0^t g^s(t-s) \triangle \overline{K}(s)\,ds. \tag{4.8}$$

If $G^s(t) > 0$ then, along with the fact that $\triangle \overline{K}(s) < \delta$ for all $s \in [0,t)$, this implies $\triangle \overline{K}(t) < \delta G^s(t) \leq \delta$. On the other hand, if $G^s(t) = 0$, it must be that $g^s(s) = 0$ for a.e. $s \in [0,t]$ and so (4.8) implies that $\triangle \overline{K}(t) = 0 \leq \delta$. Thus (4.7) follows in either case. In addition, the right-continuity of $\overline{K}^1$ and $\overline{K}^2$ implies that $\triangle \overline{K}(\tau) \geq \delta$. When combined with (4.7), (3.12) and (3.16), this shows that

$$\overline{X}^1(\tau) = \langle \mathbf{1}, \overline{\nu}_\tau^1 \rangle < 1 \quad \text{and} \quad \overline{Q}^1(\tau) = 0. \tag{4.9}$$

Now, define

$$r \doteq \sup\{t < \tau : \overline{Q}^2(t) < \overline{Q}^1(t)\} \vee 0.$$

Then for every $t \in [r, \tau]$, $\overline{Q}^2(t) \geq \overline{Q}^1(t)$. If $r = 0$, then $\triangle\overline{K}(r) = \triangle\overline{K}(0) = 0 < \delta$. On the other hand, if $r > 0$, there exists a sequence of $\{t_n\}_{n=1}^{\infty}$ such that $t_n < r$ and $t_n \to r$ as $n \to \infty$ and $0 \leq \overline{Q}^2(t_n) < \overline{Q}^1(t_n)$ for each $n \in \mathbb{N}$. Since $\overline{Q}^1$ and $\overline{Q}^2$ are càdlàg, this implies that

$$\overline{Q}^2(r-) \leq \overline{Q}^1(r-), \tag{4.10}$$

and, due to (3.12) and (3.16), it also follows that $\overline{X}^1(t_n) > \langle \mathbf{1}, \overline{\nu}^1_{t_n} \rangle = 1$ for every $n \in \mathbb{N}$. When combined with (4.8), this shows that for $n \in \mathbb{N}$,

$$\triangle\overline{K}(t_n) \leq \int_0^{t_n} g^s(t_n - s)\triangle\overline{K}(s)\,ds = \int_0^{t_n} g^s(s)\triangle\overline{K}(t_n - s)\,ds.$$

Since $\overline{K}^1$ and $\overline{K}^2$ are càdlàg, this implies that

$$\triangle\overline{K}(r-) \leq \int_0^r g^s(s)\triangle\overline{K}\big((r-s)-\big)\,ds.$$

Using the fact that $\triangle\overline{K}((r-s)-) < \delta$ for all $s \in (0, r)$, it is easy to see [once again, as in the analysis of (4.8), by considering the cases $G^s(r) > 0$ and $G^s(r) = 0$ separately] that this implies

$$\triangle\overline{K}(r-) < \delta. \tag{4.11}$$

On the other hand, since (3.18) is satisfied with $(\overline{K}, \overline{R}, \overline{Q})$ replaced by $(\overline{K}^i, \overline{R}^i, \overline{Q}^i)$ for $i = 1, 2$, and $\triangle\overline{Q}(0) + \triangle\overline{E}(t) = 0$ for each $t \geq 0$, it follows that

$$\triangle\overline{K}(\tau) + \triangle\overline{R}(\tau) + \triangle\overline{Q}(\tau) = \triangle\overline{K}(r-) + \triangle\overline{R}(r-) + \triangle\overline{Q}(r-) = 0.$$

Hence,

$$\triangle\overline{K}(\tau) - \triangle\overline{K}(r-) = -\big(\triangle\overline{R}(\tau) - \triangle\overline{R}(r-)\big) - \triangle\overline{Q}(\tau) + \triangle\overline{Q}(r-).$$

Since $-\triangle\overline{Q}(\tau) = \overline{Q}^1(\tau) - \overline{Q}^2(\tau) = -\overline{Q}^2(\tau) \leq 0$ due to (4.9) and $\triangle\overline{Q}(r-) \leq 0$ by (4.10), we obtain

$$\triangle\overline{K}(\tau) - \triangle\overline{K}(r-) \leq -\big(\triangle\overline{R}(\tau) - \triangle\overline{R}(r-)\big). \tag{4.12}$$

We now show that the right-hand side of the above display is nonpositive. For each $t \geq 0$, by (3.14), we see that

$$\begin{aligned}\triangle\overline{R}(t) &= \overline{R}^2(t) - \overline{R}^1(t)\\ &= \int_0^t \left(\int_0^{\overline{Q}^2(s)} h^r((\overline{F}^{\overline{\eta}^2_s})^{-1}(y))\,dy\right) ds\\ &\quad - \int_0^t \left(\int_0^{\overline{Q}^1(s)} h^r((\overline{F}^{\overline{\eta}^1_s})^{-1}(y))\,dy\right) ds.\end{aligned}$$

Since $\overline{\eta}^1 = \overline{\eta}^2$, it follows that $\overline{F}^{\overline{\eta}^1_\cdot} = \overline{F}^{\overline{\eta}^2_\cdot}$. Together with the continuity of $\overline{R}^1$ and $\overline{R}^2$, this yields the equation

$$
\begin{aligned}
\triangle\overline{R}(\tau) - \triangle\overline{R}(r-) \\
&= \triangle\overline{R}(\tau) - \triangle\overline{R}(r) \\
&= \int_r^\tau \left(\int_0^{\overline{Q}^2(s)} h^r((\overline{F}^{\overline{\eta}^1_s})^{-1}(y))\,dy \right) ds \\
&\quad - \int_r^\tau \left(\int_0^{\overline{Q}^1(s)} h^r((\overline{F}^{\overline{\eta}^1_s})^{-1}(y))\,dy \right) ds.
\end{aligned}
\tag{4.13}
$$

However, by the definition of r, for each $t \in [r, \tau]$, $\overline{Q}^2(t) \geq \overline{Q}^1(t)$, and so $\triangle\overline{R}(\tau) - \triangle\overline{R}(r-) \geq 0$. Together with (4.12) and (4.11), this implies

$$\triangle\overline{K}(\tau) \leq \triangle\overline{K}(r-) < \delta.$$

Essentially the same argument can be used to also show that $\triangle\overline{K}(\tau-) \leq \triangle\overline{K}(r-) < \delta$. Hence $\triangle\overline{K}(\tau) \vee \triangle\overline{K}(\tau-) < \delta$, which contradicts the definition of τ.

Thus we have proved that $\tau = \infty$ and $\overline{K}^2(t) - \overline{K}^1(t) \leq \delta$ for each $\delta > 0$ and $t \geq 0$. By letting $\delta \to 0$, we have $\overline{K}^2(t) \leq \overline{K}^1(t)$ for all $t \geq 0$. An exactly analogous argument yields the reverse inequality $\overline{K}^1(t) \leq \overline{K}^2(t)$ for each $t \geq 0$, and so it must be that $\overline{K}^2 = \overline{K}^1$. By Corollary 4.2, it follows that $\overline{\nu}^1 = \overline{\nu}^2$. Also, by (3.18), we obtain

$$\overline{R}^1 + \overline{Q}^1 = \overline{R}^2 + \overline{Q}^2. \tag{4.14}$$

We now show that, in fact $\overline{Q}^1 = \overline{Q}^2$ and $\overline{R}^1 = \overline{R}^2$. If there exists $t \in (0, \infty)$ such that $\overline{Q}^1(t) > \overline{Q}^2(t)$, let

$$s \doteq \sup\{v < t : \overline{Q}^1(v) \leq \overline{Q}^2(v)\} \vee 0.$$

Then $\overline{Q}^1(s-) \leq \overline{Q}^2(s-)$ and $\overline{Q}^1(v) > \overline{Q}^2(v)$ for each $v \in (s, t]$. Due to the fact that $\overline{\eta}^1 = \overline{\eta}^2$, we have

$$
\begin{aligned}
\overline{R}^1(t) - \overline{R}^1(s) &= \int_s^t \left(\int_0^{\overline{Q}^1(v)} h^r((\overline{F}^{\overline{\eta}^1_v})^{-1}(y))\,dy \right) dv \\
&\geq \int_s^t \left(\int_0^{\overline{Q}^2(v)} h^r((\overline{F}^{\overline{\eta}^2_v})^{-1}(y))\,dy \right) dv \\
&= \overline{R}^2(t) - \overline{R}^2(s).
\end{aligned}
$$

From (4.14) and the continuity of $\overline{R}^i$, $i = 1, 2$, we deduce that $\overline{Q}^1(t) - \overline{Q}^1(s-) \leq \overline{Q}^2(t) - \overline{Q}^2(s-)$. Combining this with the inequality $\overline{Q}^1(s-) \leq \overline{Q}^2(s-)$ proved above, we obtain $\overline{Q}^1(t) \leq \overline{Q}^2(t)$, which leads to a contradiction. Hence $\overline{Q}^1(v) \leq$

$\overline{Q}^2(v)$ for all $v \in (0, \infty)$. By symmetry, we can also argue that $\overline{Q}^1(v) \geq \overline{Q}^2(v)$ for all $v \in (0, \infty)$. This shows $\overline{Q}^1 = \overline{Q}^2$ and, hence, $\overline{R}^1 = \overline{R}^2$. Finally, by (3.12), we have $\overline{X}^1 = \overline{X}^2$. □

PROOF OF THEOREM 3.5. The first statement in Theorem 3.5 follows from Theorem 4.6. The second statement follows directly from Corollary 4.2 and the fluid equations (3.12), (3.14), (3.15) and (3.17). The alternative representation (3.24) for $\overline{K}$ is a direct consequence of the renewal equation (4.5) and the fact that the first three terms on the right-hand side of (4.5) are bounded by one. □

REMARK 4.7. For future use, we observe here that the result of Lemma 5.16 in [14] (and the analog with $\overline{\nu}$ replaced by $\overline{\eta}$), which was obtained for the model without abandonments, is also valid in the present context. This is because equations (3.20) and (3.21) of Theorem 3.5 and Corollary 4.14 of [14] show that, in the terminology of [14], $\{\overline{\eta}_s\}$ (resp., $\{\overline{\nu}_s\}$) satisfies the simplified age equation associated with a certain Radon measure $\xi(\overline{\eta}_0, \overline{E})$ and h^r [resp., $\xi(\overline{\nu}_0, \overline{K})$ and h^s]. Therefore, by Proposition 4.15 of [14], it follows that the result of Lemma 5.16 of [14] is also valid in the present context.

5. A family of martingales.

In Section 5.1, we identify the compensators (with respect to the filtration $\mathcal{F}_t^{(N)}$) of the cumulative departure, potential reneging and (actual) reneging processes. Then, in Section 5.2, we establish a more convenient representation for the compensator of the reneging process.

5.1. *Compensators.*

For any bounded measurable function φ on $[0, H^s) \times \mathbb{R}_+$, consider the sequence $\{A_{\varphi,\nu}^{(N)}\}$ of processes given by

$$A_{\varphi,\nu}^{(N)}(t) \doteq \int_0^t \left(\int_{[0,H^s)} \varphi(x,s) h^s(x) \nu_s^{(N)}(dx) \right) ds, \qquad t \in [0, \infty). \tag{5.1}$$

Likewise, for any bounded measurable function φ on $[0, H^r) \times \mathbb{R}_+$ and $N \in \mathbb{N}$, let

$$A_{\varphi,\eta}^{(N)}(t) \doteq \int_0^t \left(\int_{[0,H^r)} \varphi(x,s) h^r(x) \eta_s^{(N)}(dx) \right) ds, \qquad t \in [0, \infty). \tag{5.2}$$

In Proposition 5.1, we show that $A_{\varphi,\nu}^{(N)}$ (resp., $A_{\varphi,\eta}^{(N)}$) is the $\mathcal{F}_t^{(N)}$-compensator of the associated "φ-weighted" cumulative departure process $D_\varphi^{(N)}$ (resp., $S_\varphi^{(N)}$). A similar result was established in [14] for the model without abandonments. However, the filtration $\{\mathcal{F}_t^{(N)}\}$ considered here is larger than the one considered in [14], and so Proposition 5.1 does not directly follow from the results in [14].

PROPOSITION 5.1. *The following properties hold*:

1. *For every bounded measurable function* φ *on* $[0,H^s)\times\mathbb{R}_+$ *such that the function* $s\mapsto\varphi(a_j^{(N)}(s),s)$ *is left continuous on* $[0,\infty)$ *for each* j, *the process* $M_{\varphi,\nu}^{(N)}$ *defined by*

$$M_{\varphi,\nu}^{(N)} \doteq D_\varphi^{(N)} - A_{\varphi,\nu}^{(N)} \tag{5.3}$$

is a local $\mathcal{F}_t^{(N)}$*-martingale. Moreover, for every* $N\in\mathbb{N}$, $t\in[0,\infty)$ *and* $m\in[0,H^s)$,

$$\left|A_{\varphi,\nu}^{(N)}(t)\right| \le \|\varphi\|_\infty\left(X^{(N)}(0)+E^{(N)}(t)\right)\left(\int_0^m h^s(x)\,dx\right) < \infty \tag{5.4}$$

for every $\varphi\in\mathcal{C}_c([0,H^s)\times\mathbb{R}_+)$ *with* $\operatorname{supp}(\varphi)\subset[0,m]\times\mathbb{R}_+$. *In addition, the quadratic variation process* $\langle\overline{M}_{\varphi,\nu}^{(N)}\rangle$ *of the scaled process* $\overline{M}_{\varphi,\nu}^{(N)}\doteq M_{\varphi,\nu}^{(N)}/N$ *satisfies*

$$\lim_{N\to\infty}\mathbb{E}\left[\langle\overline{M}_{\varphi,\nu}^{(N)}\rangle(t)\right]=0;\qquad \overline{M}_{\varphi,\nu}^{(N)}\Rightarrow\mathbf{0}\qquad \text{as } N\to\infty. \tag{5.5}$$

2. *Furthermore, properties* (5.3)–(5.5) *also hold with* D, a_j, ν, H^s *and* h^s, *respectively, replaced by* S, w_j, η, H^r *and* h^r.

PROOF. In Lemma 5.4 and Corollary 5.5 of [14], it was shown that $A_{\varphi,\nu}^{(N)}$ is the compensator of $D_\varphi^{(N)}$ with respect to a certain filtration. The filtration $\{\mathcal{F}_t^{(N)}\}$ that we consider here is larger than the filtration used in [14] since it also includes the σ-algebra generated by the potential waiting times $\{\eta_j^{(N)}(s), s\le t, j=-\mathcal{E}_0^{(N)}+1,\dots,E^{(N)}(t)\}$. Thus the results of [14] do not directly apply here. Nevertheless, as we prove below, the result continues to hold due to the assumed independence of the patience and service times.

We first claim that for every $\mathcal{F}_t^{(N)}$-stopping time Υ,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\{\theta_n^k\le j/2^m<\Upsilon,\zeta_n^k>j/2^m\}}\mathbb{1}_{\{\zeta_n^k\le(j+1)/2^m\}}|\mathcal{F}_{j/2^m}^{(N)}\right]\\&\qquad=\mathbb{1}_{\{\theta_n^k\le j/2^m<\Upsilon,\zeta_n^k>j/2^m\}}\int_{j/2^m}^{(j+1)/2^m}\frac{g^s(u-\theta_n^k)}{1-G^s(j/2^m-\theta_n^k)}\,du,\end{aligned} \tag{5.6}$$

where θ_n^k (resp., ζ_n^k) is the time at which the nth customer to be served at station k starts (resp., completes) service. Then $\zeta_n^k-\theta_n^k$ is the service time of the nth customer to be served at station k, which has cumulative distribution function G^s. In order to show the equality in (5.6), it suffices to show that for every bounded $\mathcal{F}_{j/2^m}^{(N)}$-adapted random variable H,

$$\begin{aligned}&\mathbb{E}\left[H\mathbb{1}_{\{\theta_n^k\le j/2^m<\Upsilon,\zeta_n^k>j/2^m\}}\mathbb{1}_{\{\zeta_n^k\le(j+1)/2^m\}}\right]\\&\qquad=\mathbb{E}\left[H\mathbb{1}_{\{\theta_n^k\le j/2^m<\Upsilon,\zeta_n^k>j/2^m\}}\int_{j/2^m}^{(j+1)/2^m}\frac{g^s(u-\theta_n^k)}{1-G^s(j/2^m-\theta_n^k)}\,du\right].\end{aligned} \tag{5.7}$$

For $j \in \mathbb{N}$, $m \in \mathbb{N}$, define $\mathcal{G}^{(N)}_{j/2^m}$ be the σ-algebra to be generated by the events $\{(\theta^k_n \le x) \cap (\theta^k_n \le \frac{j}{2^m}, \zeta^k_n > \frac{j}{2^m}), x \ge 0\}$. In particular, $\mathcal{G}^{(N)}_{j/2^m}$ contains the information of the ages of all customers in service at time $\frac{j}{2^m}$. Recall that the patience times and the service times of customers are assumed to be independent. Therefore, given $\mathcal{G}^{(N)}_{j/2^m}$, $\zeta^k_n - \theta^k_n$ and $\mathcal{F}^{(N)}_{j/2^m}$ are conditionally independent. Hence, it follows from the left-hand side of (5.7) that

$$\begin{aligned}
&\mathbb{E}\big[H\mathbb{1}_{\{\theta^k_n \le j/2^m < \Upsilon, \zeta^k_n > j/2^m\}}\mathbb{1}_{\{\zeta^k_n \le (j+1)/2^m\}}\big] \\
&\quad = \mathbb{E}\big[\mathbb{E}\big[H\mathbb{1}_{\{j/2^m < \Upsilon\}}\mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}}\mathbb{1}_{\{\zeta^k_n - \theta^k_n \le (j+1)/2^m - \theta^k_n\}} | \mathcal{G}^{(N)}_{j/2^m}\big]\big] \\
&\quad = \mathbb{E}\big[\mathbb{E}\big[H\mathbb{1}_{\{j/2^m < \Upsilon\}} | \mathcal{G}^{(N)}_{j/2^m}\big] \\
&\qquad \times \mathbb{E}\big[\mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}}\mathbb{1}_{\{\zeta^k_n - \theta^k_n \le (j+1)/2^m - \theta^k_n\}} | \mathcal{G}^{(N)}_{j/2^m}\big]\big]
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}}\mathbb{1}_{\{\zeta^k_n - \theta^k_n \le (j+1)/2^m - \theta^k_n\}} | \mathcal{G}^{(N)}_{j/2^m}\big] \\
&\quad = \mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}} \int_{j/2^m}^{(j+1)/2^m} \frac{g^s(u - \theta^k_n)}{1 - G^s(j/2^m - \theta^k_n)}\, du.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{E}\big[H\mathbb{1}_{\{j/2^m < \Upsilon\}} | \mathcal{G}^{(N)}_{j/2^m}\big]\mathbb{E}\big[\mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}}\mathbb{1}_{\{\zeta^k_n - \theta^k_n \le (j+1)/2^m - \theta^k_n\}} | \mathcal{G}^{(N)}_{j/2^m}\big]\big] \\
&\quad = \mathbb{E}\bigg[\mathbb{E}\big[H\mathbb{1}_{\{j/2^m < \Upsilon\}} | \mathcal{G}^{(N)}_{j/2^m}\big]\mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}} \\
&\qquad\qquad \times \int_{j/2^m}^{(j+1)/2^m} \frac{g^s(u - \theta^k_n)}{1 - G^s(j/2^m - \theta^k_n)}\, du\bigg] \\
&\quad = \mathbb{E}\bigg[\mathbb{E}\bigg[H\mathbb{1}_{\{j/2^m < \Upsilon\}}\mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}} \\
&\qquad\quad \times \int_{j/2^m}^{(j+1)/2^m} \frac{g^s(u - \theta^k_n)}{1 - G^s(j/2^m - \theta^k_n)}\, du \bigg| \mathcal{G}^{(N)}_{j/2^m}\bigg]\bigg] \\
&\quad = \mathbb{E}\bigg[H\mathbb{1}_{\{j/2^m < \Upsilon\}}\mathbb{1}_{\{\theta^k_n \le j/2^m, \zeta^k_n > j/2^m\}} \int_{j/2^m}^{(j+1)/2^m} \frac{g^s(u - \theta^k_n)}{1 - G^s(j/2^m - \theta^k_n)}\, du\bigg].
\end{aligned}$$

This shows that (5.7), and therefore (5.6), holds.

If φ is bounded, measurable and such that the function $s \mapsto \varphi(a^{(N)}_j(s), s)$ is left continuous for each j, then the process $\{\varphi(a^{(N)}_j(s), s), s \ge 0\}$ is $\mathcal{F}^{(N)}_t$-predictable. Therefore, it follows from the standard theory (cf. Theorem 3.18 of [10]) that $M^{(N)}_{\varphi,\nu}$ is a local $\mathcal{F}^{(N)}_t$-martingale. Inequality (5.4) can be established exactly as in Proposition 5.7 of [14] and assertions (5.5) can be proved using the same argument as in

Lemma 5.9 of [14], thus establishing property (1). Due to the analogy between the service dynamics and the potential queue dynamics (see Remark 2.2), property (2) is a direct consequence of property (1). □

REMARK 5.2. It is easy to see that Lemmas 5.6 and 5.8 of [14] continue to be valid in the presence of abandonments. Indeed, the proofs of Lemmas 5.6 and 5.8 of [14] only depend on Assumption 1 and Corollary 5.5 therein (since, as shown in Lemma 5.12 of [14], the additional conditions (5.32) and (5.33) of Lemma 5.8 of [14] can be derived from Assumption 1), which correspond to Assumption 3.1 and Proposition 5.1 of this paper. In addition, due to the parallels between the dynamics of $\nu^{(N)}$ and $\eta^{(N)}$ (see Remark 2.2), the analogs of the results in Lemmas 5.6 and 5.8, with $D^{(N)}$, $\nu^{(N)}$, G^s and H^s, respectively, replaced by $S^{(N)}$, $\eta^{(N)}$, G^r and H^r, also hold. In this case, even though $\eta_0^{(N)}$ is (unlike $\nu_0^{(N)}$) not necessarily a sub-probability measure, the verification of the conditions analogous to (5.32) and (5.33) of Lemma 5.8 in [14] can still be carried out in the same manner since Assumption 3.1 implies that the sequence $\{\langle \mathbf{1}, \eta_0^{(N)}\rangle\}$ is tight. Moreover, even though G^r is allowed to have a mass at ∞, the proofs of Lemmas 5.6 and 5.8 are still valid, with the renewal function U^s now replaced by the function $U^r(\cdot) = \int_0^{\cdot} \sum_{n=1}^{\infty} (g^r)^{*n}(s)\,ds$, where $(g^r)^{*n}$ is the nth convolution of g^r on $[0, \infty)$.

Now, note from (2.25) that $R^{(N)} = S^{(N)}_{\theta^{(N)}}$, where $\theta^{(N)}$ is defined by (2.26). In view of the fact that $A^{(N)}_{\varphi,\eta}$ is the compensator for $S^{(N)}_{\varphi}$, it is natural to conjecture that the compensator of $R^{(N)}$ is equal to $A^{(N)}_{\theta^{(N)},\eta}$, where

$$A^{(N)}_{\theta^{(N)},\eta}(t) \doteq \int_0^t \left(\int_{[0,H^r)} \mathbb{1}_{[0,\chi^{(N)}(s-)]}(x) h^r(x) \eta_s^{(N)}(dx) \right) ds, \qquad t \in [0, \infty). \tag{5.8}$$

However, this is not immediate from Proposition 5.1(2) since $\theta^{(N)}(w_j^{(N)}(\cdot), \cdot)$ is not left continuous for any j. Instead, we approximate $\theta^{(N)}$ by a sequence $\{\theta_m^{(N)}\}_{N\in\mathbb{N}}$ defined by

$$\theta_m^{(N)}(x, s) \doteq \mathbb{1}_{(x-1/m,\infty)}(\chi^{(N)}(s-)), \tag{5.9}$$

which is shown to be left continuous in Lemma 5.3. Then in Lemma 5.4, we use an approximation argument to show that $A^{(N)}_{\theta^{(N)},\eta}$ is indeed the compensator of $R^{(N)}$.

LEMMA 5.3. *For each $m \geq 1$, $x \in \mathbb{R}$ and $s \in \mathbb{R}_+$, the sequence $\{\theta_m^{(N)}\}_{N\in\mathbb{N}}$ defined by (5.9) satisfies the following two properties*:

1. *For every $N \in \mathbb{N}$, $x \in \mathbb{R}$, $s \in \mathbb{R}$, $\theta_m^{(N)}(x, s)$ is nonincreasing in m and converges, as $m \to \infty$, to $\theta^{(N)}(x, s)$ for every sample point in Ω.*

2. *For each* $N, m \in \mathbb{R}$, $j \in \mathbb{Z}$, *the process* $\theta_m^{(N)}(w_j^{(N)}(\cdot), \cdot)$ *has left continuous paths on* $(0, \infty)$.

PROOF. The first property is immediate from the definition of $\theta_m^{(N)}$. For the second property, fix $N, m \in \mathbb{N}$, $s > 0$, $j \in \mathbb{Z}$ and $\omega \in \Omega$. To ease the notation, we shall suppress ω from the notation. Let $\{s_n\}$ be a sequence in $(0, \infty)$ such that $s_n \uparrow s$ as $n \to \infty$. We now consider two mutually exclusive cases.

Case 1. $\theta_m^{(N)}(w_j^{(N)}(s), s) = 1$. Then $w_j^{(N)}(s) < \chi^{(N)}(s-) + 1/m$. Since $w_j^{(N)}$ is nondecreasing, $w_j^{(N)}(s_n) \le w_j^{(N)}(s)$ and since the process $\{\chi^{(N)}(s-), s \ge 0\}$ is left continuous, we have, for all n large enough, $w_j^{(N)}(s_n) < \chi^{(N)}(s_n-) + 1/m$. Hence, $\theta_m^{(N)}(w_j^{(N)}(s_n), s_n) = 1$ for all $n \in \mathbb{N}$. Thus, in this case, $\theta_m^{(N)}(w_j^{(N)}(\cdot), \cdot)$ is left continuous at s.

Case 2. $\theta_m^{(N)}(w_j^{(N)}(s), s) = 0$. Then $w_j^{(N)}(s) \ge \chi^{(N)}(s-) + 1/m$. It follows from Lemma A.2 that for all sufficiently large n, $\chi^{(N)}(s-) - \chi^{(N)}(s_n-) = s - s_n > 0$. Since (2.2) implies $w_j^{(N)}(s) - w_j^{(N)}(s_n) \le s - s_n$ for all $n \in \mathbb{N}$, this implies $w_j^{(N)}(s_n) \ge \chi^{(N)}(s_n-) + 1/m$ for all n large enough. Hence, $\theta_m^{(N)}(w_j^{(N)}(s_n), s_n) = 0$ and $\theta_m^{(N)}(w_j^{(N)}(\cdot), \cdot)$ is again left continuous at s. $\square$

LEMMA 5.4. *For every* $N \in \mathbb{N}$, *the process* $M_{\theta^{(N)},\eta}^{(N)}$ *defined by*

$$M_{\theta^{(N)},\eta}^{(N)} \doteq R^{(N)} - A_{\theta^{(N)},\eta}^{(N)} \tag{5.10}$$

is a local $\mathcal{F}_t^{(N)}$*-martingale. In addition, as* $N \to \infty$,

$$\mathbb{E}[\langle \overline{M}_{\theta^{(N)},\eta}^{(N)} \rangle(t)] \to 0, \qquad \overline{M}_{\psi,\eta}^{(N)} \Rightarrow \mathbf{0} \quad \textit{and} \quad \overline{M}_{\theta^{(N)},\eta}^{(N)} \Rightarrow \mathbf{0}. \tag{5.11}$$

PROOF. Fix $N \in \mathbb{N}$, and let $A_{\theta_m^{(N)},\eta}^{(N)}$, $m \in \mathbb{N}$, be defined in the obvious way

$$A_{\theta_m^{(N)},\eta}^{(N)}(t) \doteq \int_0^t \left(\int_{[0,H^r)} \theta_m^{(N)}(x, s) h^r(x) \eta_s^{(N)}(dx) \right) ds. \tag{5.12}$$

By Proposition 5.1(2) and Lemma 5.3(2), the process $A_{\theta_m^{(N)},\eta}^{(N)}$ is the $\mathcal{F}_t^{(N)}$-compensator of the process $S_{\theta_m^{(N)}}^{(N)}$, and the process $M_{\theta_m^{(N)},\eta}^{(N)}$ defined by

$$M_{\theta_m^{(N)},\eta}^{(N)} \doteq S_{\theta_m^{(N)}}^{(N)} - A_{\theta_m^{(N)},\eta}^{(N)} \tag{5.13}$$

is a local $\mathcal{F}_t^{(N)}$-martingale. Now, by Lemma 5.3(1), $\theta_m^{(N)} \to \theta^{(N)}$ pointwise on $\mathbb{R}_+^2$, $|\theta_m^{(N)}(x, s) - \theta^{(N)}(x, s)| \le 1$ for all $(x, s) \in \mathbb{R}_+^2$, and $\mathbb{E}[S_{\mathbf{1}}^{(N)}(t)] < \infty$,

$\mathbb{E}[A^{(N)}_{\mathbf{1},\eta}(t)] < \infty$ for all $t \in (0,\infty)$. Hence, an application of the dominated convergence theorem shows that for all $t \in (0,\infty)$, as $m \to \infty$,

$$\mathbb{E}\Big[\sup_{0\le s\le t} \big|A^{(N)}_{\theta^{(N)}_m,\eta}(s) - A^{(N)}_{\theta^{(N)},\eta}(s)\big|\Big] \to 0$$

and

$$\mathbb{E}\Big[\sup_{0\le s\le t} \big|S^{(N)}_{\theta^{(N)}_m}(s) - S^{(N)}_{\theta^{(N)}}(s)\big|\Big] \to 0,$$

and hence $M^{(N)}_{\theta^{(N)}_m,\eta}$ converges in distribution to $M^{(N)}_{\theta^{(N)},\eta}$. Since $|S^{(N)}_{\theta^{(N)}_m}(t) - S^{(N)}_{\theta^{(N)}_m}(t-)| \le 1$ for all $t \in [0,\infty)$ and $m \in \mathbb{N}$, we conclude that $M^{(N)}_{\theta^{(N)},\eta}$ is a local $\mathcal{F}^{(N)}_t$-martingale by Corollary 1.19 of Chapter IX of [10]. Given that $M^{(N)}_{\theta^{(N)},\eta}$ is a martingale, the proof of the limits (5.11) is exactly analogous to the proof of (5.5), as carried out in Lemma 5.9 of [14]. □

5.2. *An alternative representation for the compensator of* $R^{(N)}$. We now derive an alternative, more convenient representation for $A^{(N)}_{\theta^{(N)},\eta}$, or more generally, for processes of the form $A^{(N)}_{\theta^{(N)},\eta}$, but with h^r replaced by an arbitrary measurable function h. In what follows, recall that $F^{\eta^{(N)}_t}(x) = \eta^{(N)}_t[0,x]$ and its inverse $(F^{\eta^{(N)}_t})^{-1}$ is as defined in (1.1).

PROPOSITION 5.5. *For each* $N \in \mathbb{N}$, $t \ge 0$ *and measurable function* h *on* $[0,H^r)$,

$$\int_{[0,H^r)} \mathbb{1}_{[0,\chi^{(N)}(t-)]}(x)h(x)\eta^{(N)}_t(dx) = \int_0^{Q^{(N)}(t)+\iota^{(N)}(t)} h((F^{\eta^{(N)}_t})^{-1}(y))\,dy, \tag{5.14}$$

where

$$\iota^{(N)}(t) \doteq \begin{cases} 0, & \text{if } (\chi^{(N)}(t-) - \chi^{(N)}(t))(K^{(N)}(t) - K^{(N)}(t-)) = 0, \\ 1, & \text{if } (\chi^{(N)}(t-) - \chi^{(N)}(t))(K^{(N)}(t) - K^{(N)}(t-)) > 0. \end{cases} \tag{5.15}$$

PROOF. Fix $N \in \mathbb{N}$, $t \ge 0$ and a measurable function h on $[0,H^r)$. By the representation (2.3) for $\eta^{(N)}_t$, we have

$$\int_{[0,H^r)} \mathbb{1}_{[0,\chi^{(N)}(t-)]}(x)h(x)\eta^{(N)}_t(dx) = \sum_{j=-\mathcal{E}^{(N)}_0+1}^{E^{(N)}(t)} h(w^{(N)}_j(t))\mathbb{1}_{\{w^{(N)}_j(t)\le\chi^{(N)}(t-)\}}\mathbb{1}_{\{w^{(N)}_j(t)<r_j\}}. \tag{5.16}$$

Moreover, by (2.6),

$$Q^{(N)}(t) = \eta_t^{(N)}[0, \chi^{(N)}(t)] = \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \mathbb{1}_{\{w_j^{(N)}(t)\le \chi^{(N)}(t)\}} \mathbb{1}_{\{w_j^{(N)}(t)<r_j\}}.$$

Thus $Q^{(N)}(t)$ is the total number of customers who have arrived to the system and have not reneged by t and whose potential waiting times at t are less than or equal to $\chi^{(N)}(t)$. If we arrange those customers in increasing order of their potential waiting times at t, then for $i = 1, 2, \dots, Q^{(N)}(t)$, $(F^{\eta_t^{(N)}})^{-1}(i)$ is exactly the potential waiting time at t of the ith customer from the back of the queue.

Suppose that $(\chi^{(N)}(t-) - \chi^{(N)}(t))(K^{(N)}(t) - K^{(N)}(t-)) = 0$. This implies that either $\chi^{(N)}(t-) = \chi^{(N)}(t)$ holds or both $\chi^{(N)}(t-) > \chi^{(N)}(t)$ and $K^{(N)}(t) = K^{(N)}(t-)$ hold. The latter condition indicates that the head-of-the-line customer right before time t reneged at time t. In this case, the right-hand side of (5.16) admits the alternative representation

$$\int_0^{Q^{(N)}(t)} h((F^{\eta_t^{(N)}})^{-1}(y))\, dy.$$

On the other hand, suppose that $(\chi^{(N)}(t-) - \chi^{(N)}(t))(K^{(N)}(t) - K^{(N)}(t-)) > 0$. In this case, the head-of-the-line customer right before time t departs for service at time t and this customer is counted in the right-hand side of (5.16) but not in $Q^{(N)}(t)$. Since $E^{(N)}(t) - E^{(N)}(t-) \le 1$, there is exactly one such customer, that is, $K^{(N)}(t) - K^{(N)}(t-) = 1$. Hence the right-hand side of (5.16) can be rewritten as

$$\int_0^{Q^{(N)}(t)+1} h((F^{\eta_t^{(N)}})^{-1}(y))\, dy. \qquad \square$$

As an immediate consequence of (5.8), Lemma 5.4, and Proposition 5.5, we obtain the following alternative representation for the compensator $A^{(N)}_{\theta^{(N)},\eta}$ of $R^{(N)}$:

$$A^{(N)}_{\theta^{(N)},\eta}(t) \doteq \int_0^t \left(\int_0^{Q^{(N)}(t)+\iota^{(N)}(t)} h^r((F^{\eta_s^{(N)}})^{-1}(y))\, dy \right) ds, \qquad t \in [0, \infty),$$

where $\iota^{(N)}$ is given by (5.15).

6. Tightness of pre-limit sequence. The main objective of this section is to show that, under suitable assumptions, the sequence of scaled state processes $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})\}$ and the sequences of auxiliary processes are tight. Specifically, from (2.23) and (5.4) it is clear that for every t, the linear functionals $\overline{D}^{(N)}_{\cdot}(t) : \varphi \mapsto \overline{D}^{(N)}_{\varphi}(t)$ and $\overline{A}^{(N)}_{\cdot,\nu}(t) : \varphi \mapsto \overline{A}^{(N)}_{\varphi,\nu}(t)$ are finite Radon measures on

$[0, H^s) \times \mathbb{R}_+$. Likewise, from (2.24) and the fact that (5.4) holds with ν, h^s, respectively, replaced by η, h^r by property (2) of Proposition 5.1, it follows that the linear functionals $\overline{S}_\cdot^{(N)}(t) : \psi \mapsto \overline{S}_\psi^{(N)}(t)$ and $\overline{A}_{\cdot,\eta}^{(N)}(t) : \psi \mapsto \overline{A}_{\psi,\eta}^{(N)}(t)$ define finite Radon measures on $[0, H^r) \times \mathbb{R}_+$. Thus $\{\overline{D}_\cdot^{(N)}(t) : t \in [0,\infty)\}$ and $\{\overline{A}_{\cdot,\nu}^{(N)}(t) : t \in [0,\infty)\}$ can be viewed as $\mathcal{M}_F([0, H^s) \times \mathbb{R}_+)$-valued càdlàg processes, and $\{\overline{S}_\cdot^{(N)}(t) : t \in [0,\infty)\}$ and $\{\overline{A}_{\cdot,\eta}^{(N)}(t) : t \in [0,\infty)\}$ can be viewed as $\mathcal{M}_F([0, H^r) \times \mathbb{R}_+)$-valued càdlàg processes. Now, for $N \in \mathbb{N}$, let

$$\begin{aligned} \overline{Y}^{(N)} \doteq \big(&\overline{X}^{(N)}(0), \overline{E}^{(N)}, \overline{X}^{(N)}, \overline{R}^{(N)}, \overline{\nu}_0^{(N)}, \\ &\overline{\nu}^{(N)}, \overline{\eta}_0^{(N)}, \overline{\eta}^{(N)}, \overline{A}_{\cdot,\nu}^{(N)}, \overline{D}_\cdot^{(N)}, \overline{A}_{\cdot,\eta}^{(N)}, \overline{S}_\cdot^{(N)}\big). \end{aligned} \tag{6.1}$$

Then each $\overline{Y}^{(N)}$ is a $\mathcal{Y}$-valued process, where $\mathcal{Y}$ is the space

$$\begin{aligned} \mathcal{Y} \doteq \mathbb{R}_+ &\times (\mathcal{D}_{\mathbb{R}_+}[0,\infty))^3 \times \mathcal{M}_F[0, H^s) \times \mathcal{D}_{\mathcal{M}_F[0,H^s)}[0,\infty) \times \mathcal{M}_F[0, H^r) \\ &\times \mathcal{D}_{\mathcal{M}_F[0,H^r)}[0,\infty) \times \big(\mathcal{D}_{\mathcal{M}_F([0,H^s)\times\mathbb{R}_+)}[0,\infty)\big)^2 \\ &\times \big(\mathcal{D}_{\mathcal{M}_F([0,H^r)\mathbb{R}_+)}[0,\infty)\big)^2 \end{aligned}$$

equipped with the product metric. Clearly, $\mathcal{Y}$ is a Polish space. Now we state the main result of this section.

THEOREM 6.1. *Suppose Assumption 3.1 is satisfied. Then the sequence $\{\overline{Y}^{(N)}\}$ defined in (6.1) is relatively compact in the Polish space $\mathcal{Y}$, and is therefore tight.*

The relative compactness of $\{\overline{Y}^{(N)}\}$ follows from Assumption 3.1 and Lemmas 6.3, 6.4, 6.6 and 6.7 below. Since $\mathcal{Y}$ is a Polish space, tightness is then a direct consequence of Prohorov's theorem.

We start by recalling Kurtz's criteria (see Theorem 3.8.6 of [6] for details) for the relative compactness of a sequence $\{\overline{F}^{(N)}\}$ of processes in $\mathcal{D}_{\mathbb{R}_+}[0,\infty)$.

PROPOSITION 6.2 (Kurtz's criteria). *The sequence of processes $\{\overline{\mathcal{Z}}^{(N)}\}$ is relatively compact if and only if the following two properties hold*:

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

$$\lim_{R\to\infty} \sup_N \mathbb{P}\big(\overline{\mathcal{Z}}^{(N)}(t) > R\big) = 0.$$

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

$$\lim_{\delta\to 0} \sup_N \mathbb{E}\big[\big|\overline{\mathcal{Z}}^{(N)}(t+\delta) - \overline{\mathcal{Z}}^{(N)}(t)\big|^\beta\big] = 0. \tag{6.2}$$

LEMMA 6.3. *Suppose Assumption 3.1 holds. Then the sequences* $\{\overline{X}^{(N)}\}$, $\{\overline{K}^{(N)}\}$, $\{\overline{R}^{(N)}\}$, $\{\langle \mathbf{1}, \overline{\nu}^{(N)}\rangle\}$, $\{\langle \mathbf{1}, \overline{\eta}^{(N)}\rangle\}$, *the sequences* $\{\overline{D}_\varphi^{(N)}\}$, $\{\overline{A}_{\varphi,\nu}^{(N)}\}$, *for every* $\varphi \in \mathcal{C}_b([0, H^s) \times \mathbb{R}_+)$ *and the sequences* $\{S_\psi^{(N)}\}$, $\{\overline{A}_{\psi,\eta}^{(N)}\}$, *for every* $\psi \in \mathcal{C}_b([0, H^r) \times \mathbb{R}_+)$, *are relatively compact.*

PROOF. Fix $T \in (0, \infty)$. It follows from Proposition 5.1(1), (2.23) and (3.5) that for $\varphi \in \mathcal{C}_b([0, H^s) \times \mathbb{R}_+)$,

$$\sup_N \mathbb{E}\big[\overline{A}_{\varphi,\nu}^{(N)}(T)\big] = \sup_N \mathbb{E}\big[\overline{D}_\varphi^{(N)}(T)\big] \le \|\varphi\|_\infty \sup_N \mathbb{E}\big[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(T)\big] < \infty.$$

Similarly, by Proposition 5.1(2), (2.24) and (3.5), we have for every $\psi \in \mathcal{C}_b([0, H^r) \times \mathbb{R}_+)$,

$$\sup_N \mathbb{E}\big[\overline{A}_{\psi,\eta}^{(N)}(T)\big] = \sup_N \mathbb{E}\big[\overline{S}_\psi^{(N)}(T)\big] \le \|\psi\|_\infty \sup_N \mathbb{E}\big[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(T)\big] < \infty,$$

which verifies condition K1 for $\mathcal{Z} = A_{\varphi,\nu}^{(N)}, D_\varphi^{(N)}, \varphi \in \mathcal{C}_b([0, H^s) \times \mathbb{R}_+)$ and $\mathcal{Z} = A_{\psi,\eta}^{(N)}, S_\psi^{(N)}, \psi \in \mathcal{C}_b([0, H^r) \times \mathbb{R}_+)$. The same argument that was used to prove Lemma 5.8(2) in [14] can then be used to show that (6.2) is also satisfied by the same collection of $\mathcal{Z}$ (see Remark 5.2). The fact that $\overline{R}^{(N)}$ and its increments are dominated, respectively, by $\overline{S}^{(N)}$ and its increments shows that the sequence $\{\overline{R}^{(N)}\}$ also satisfies conditions K1 and K2, and is thus relatively compact. Since $\overline{D}^{(N)} = \overline{D}_{\mathbf{1}}^{(N)}$ and $\overline{S}^{(N)} = \overline{S}_{\mathbf{1}}^{(N)}$, it follows that the sequences $\{\overline{D}^{(N)}\}$ and $\{\overline{S}^{(N)}\}$ are also relatively compact. By Assumption 3.1, the sequences $\{\overline{E}^{(N)}\}$ and $\{\overline{X}^{(N)}(0)\}$ are relatively compact.

Since for every $t \ge 0$, $\langle \mathbf{1}, \overline{\nu}_t^{(N)}\rangle \le \overline{X}^{(N)}(t) \le \overline{X}^{(N)}(0) + \overline{E}^{(N)}(t)$ by (2.30) and (2.12), it follows from Markov's inequality that $\langle \mathbf{1}, \overline{\nu}_t^{(N)}\rangle$ and $\overline{X}^{(N)}$ satisfy K1 of Proposition 6.2. In addition, (2.12) also shows that

$$\begin{aligned}\big|\overline{X}^{(N)}(t) - \overline{X}^{(N)}(s)\big| &\le \big|\overline{E}^{(N)}(t) - \overline{E}^{(N)}(s)\big| + \big|\overline{D}^{(N)}(t) - \overline{D}^{(N)}(s)\big| \\ &\quad + \big|\overline{R}^{(N)}(t) - \overline{R}^{(N)}(s)\big|,\end{aligned}$$

and by (2.30) and the Lipschitz continuity of the function $x \mapsto [1 - x]^+$ with Lipschitz constant 1, we have

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

When combined with the properties of $\overline{E}^{(N)}$, $\overline{D}^{(N)}$ and $\overline{R}^{(N)}$ established above, this shows that $\{\overline{X}^{(N)}\}$ and $\{\langle \mathbf{1}, \overline{\nu}^{(N)}\rangle\}$ satisfy K2 of Proposition 6.2 and, are relatively compact. In turn, by (2.16), the relative compactness of $\{\overline{D}^{(N)}\}$ and $\{\langle \mathbf{1}, \overline{\nu}^{(N)}\rangle\}$ implies that of $\{\overline{K}^{(N)}\}$. Moreover, due to (2.13), for every $s, t \in [0, \infty)$, we have that

$$\big|\langle \mathbf{1}, \overline{\eta}_t^{(N)}\rangle - \langle \mathbf{1}, \overline{\eta}_s^{(N)}\rangle\big| \le \big|\overline{E}^{(N)}(t) - \overline{E}^{(N)}(s)\big| + \big|\overline{S}^{(N)}(t) - \overline{S}^{(N)}(s)\big|, \tag{6.3}$$

$$\langle \mathbf{1}, \overline{\eta}_t^{(N)}\rangle \le \langle \mathbf{1}, \overline{\eta}_0^{(N)}\rangle + \overline{E}^{(N)}(t). \tag{6.4}$$

Thus $\langle \mathbf{1}, \overline{\eta}^{(N)} \rangle$ is also relatively compact, and the proof is complete. $\square$

LEMMA 6.4. *Suppose Assumption 3.1 holds. For every $f \in \mathcal{C}_c^1(\mathbb{R}_+)$, the sequences $\{\langle f, \overline{\nu}^{(N)} \rangle\}$ and $\{\langle f, \overline{\eta}^{(N)} \rangle\}$ of $\mathcal{D}_{\mathbb{R}}[0, \infty)$-valued random variables are relatively compact.*

PROOF. Fix $t \in [0, \infty)$. By (2.27) and (2.28), for every $f \in \mathcal{C}_c^1(\mathbb{R}_+)$, we have

$$\langle f, \overline{\nu}_t^{(N)} \rangle - \langle f, \overline{\nu}_0^{(N)} \rangle = \int_0^t \langle f', \overline{\nu}_s^{(N)} \rangle \, ds - \overline{D}_f^{(N)}(t) + f(0)\overline{K}^{(N)}(t)$$

and

$$\langle f, \overline{\eta}_t^{(N)} \rangle - \langle f, \overline{\eta}_0^{(N)} \rangle = \int_0^t \langle f', \overline{\eta}_s^{(N)} \rangle \, ds - \overline{S}_f^{(N)}(t) + f(0)\overline{E}^{(N)}(t).$$

Since $\{\overline{D}_f^{(N)}\}$, $\{\overline{K}^{(N)}\}$, $\{\overline{S}_f^{(N)}\}$ and $\{\overline{E}^{(N)}\}$ are relatively compact due to Lemma 6.3 and property 1 of Assumption 3.1, it suffices to show that the sequences $\{\int_0^{\cdot} \langle f', \overline{\nu}_s^{(N)} \rangle \, ds\}$ and $\{\int_0^{\cdot} \langle f', \overline{\eta}_s \rangle \, ds\}$ are tight. It follows from (6.4) that for $\delta \in (0, 1)$,

$$\begin{aligned} \left| \int_t^{t+\delta} \langle f', \overline{\eta}_s^{(N)} \rangle \, ds \right| &\le \|f'\|_\infty \int_t^{t+\delta} |\langle \mathbf{1}, \overline{\eta}_s^{(N)} \rangle| \, ds \\ &\le \|f'\|_\infty \delta \big(\langle \mathbf{1}, \overline{\eta}_0^{(N)} \rangle + \overline{E}^{(N)}(t+1) \big). \end{aligned}$$

Hence, we have

$$\mathbb{E}\left[\left| \int_t^{t+\delta} \langle f', \overline{\eta}_s^{(N)} \rangle \, ds \right| \right] \le \|f'\|_\infty \delta \sup_N \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_0^{(N)} \rangle + \overline{E}^{(N)}(t+1)]. \tag{6.5}$$

For each $t \in [0, \infty)$, by (2.3) and Assumption 3.1, it follows that

$$\sup_N \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_t^{(N)} \rangle] \le \sup_N \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_0^{(N)} \rangle + \overline{E}^{(N)}(t)] < \infty. \tag{6.6}$$

Therefore, taking the limit, as $\delta \to 0$, in (6.5) and using the last inequality in (6.6), we have

$$\lim_{\delta \to 0} \sup_N \mathbb{E}\left[\left| \int_t^{t+\delta} \langle f', \overline{\eta}_s^{(N)} \rangle \, ds \right| \right] = 0.$$

Similarly, since $\langle \mathbf{1}, \overline{\nu}_s^{(N)} \rangle \le 1$ for every $s \in [0, \infty)$ and $N \in \mathbb{N}$,

$$\lim_{\delta \to 0} \sup_N \mathbb{E}\left[\left| \int_t^{t+\delta} \langle f', \overline{\nu}_s^{(N)} \rangle \, ds \right| \right] \le \lim_{\delta \to 0} \|f'\|_\infty \delta = 0.$$

Moreover, by (6.6), we also have, for every $t \in [0, \infty)$,

$$\sup_N \mathbb{E}\left[\left| \int_0^t \langle f', \overline{\eta}_s^{(N)} \rangle \, ds \right| \right] \le \|f'\|_\infty t \sup_N \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_0^{(N)} \rangle + \overline{E}^{(N)}(t)] < \infty.$$

Similarly, we have

$$\sup_N \mathbb{E}\left[\left|\int_0^t \langle f', \overline{\nu}_s^{(N)}\rangle\, ds\right|\right] \le \sup_N \mathbb{E}\left[\int_0^t |\langle f', \overline{\nu}_s^{(N)}\rangle|\, ds\right] \le \|f'\|_\infty t < \infty.$$

This implies that $\{\int_0^\cdot \langle f', \overline{\eta}_s^{(N)}\rangle\, ds\}$ and $\{\int_0^\cdot \langle f', \overline{\nu}_s^{(N)}\rangle\, ds\}$ both satisfy criteria K1 and K2 of Proposition 6.2 and hence are relatively compact. This completes the proof of the lemma. □

Next, we show that $\{\overline{\nu}^{(N)}\}$ and $\{\overline{\eta}^{(N)}\}$ are tight, and hence are relatively compact with respect to the topology on $\mathcal{D}_{\mathcal{M}_F[0,H^s)}[0,\infty)$ and $\mathcal{D}_{\mathcal{M}_F[0,H^r)}[0,\infty)$, respectively. Since, as mentioned in Section 1.3.1, $\mathcal{M}_F[0,H^s)$ and $\mathcal{M}_F[0,H^r)$, equipped with the topology of weak convergence, are Polish spaces, we can apply Jakubowski's criteria to establish the tightness of $\{\overline{\nu}^{(N)}\}$ and $\{\overline{\eta}^{(N)}\}$. For convenience, we recall Jakubowski's criteria.

PROPOSITION 6.5 (Jakubowski). *A sequence $\{\overline{\pi}^{(N)}\}$ of $\mathcal{D}_{\mathcal{M}_F[0,H)}[0,\infty)$-valued random elements defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is tight if and only if the following two conditions hold*:

J1. *For each $T > 0$ and $0 < \delta < 1$, there are compact subsets $\tilde{C}_{T,\delta}$ of $\mathcal{M}_F[0,H)$ such that*

$$\liminf_{N\to\infty} \mathbb{P}\big(\overline{\nu}_t^{(N)} \in \tilde{C}_{T,\delta} \textit{ for all } t \in [0,T]\big) > 1-\delta.$$

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

LEMMA 6.6. *Suppose Assumption 3.1 holds. The sequences $\{\overline{\nu}^{(N)}\}$ and $\{\overline{\eta}^{(N)}\}$ are relatively compact.*

PROOF. By Remark 5.11 of [14] and Lemma 6.4, it follows that $\{\overline{\nu}^{(N)}\}$ and $\{\overline{\eta}^{(N)}\}$ satisfy Jakubowski's J2 criterion. Therefore, it suffices to show that they also satisfy Jakubowski's J1 criterion. By (2) and (3) of Assumption 3.1, for almost every $\omega \in \Omega$, $\sup_N \overline{\nu}_0^{(N)}(\omega)[0,H^s) < \infty$. By Lemma A 7.5 of [12], for every $\varepsilon > 0$, there exists $k(\omega,\varepsilon) < \infty$ such that $\sup_N \overline{\nu}_0^{(N)}(\omega)(k(\omega,\varepsilon), H^s) < \varepsilon$. The argument for tightness of $\{\overline{\nu}^{(N)}\}$ (in the absence of reneging) presented in Lemma 5.12 of [14] can be directly applied to show that $\{\overline{\nu}^{(N)}\}$ satisfies Jakubowski's J1 criterion, and hence $\{\overline{\nu}^{(N)}\}$ is tight in the presence of reneging as well. Similarly, due to (2) and (4) of Assumption 3.1, for almost every $\omega \in \Omega$, $\sup_N \overline{\eta}_0^{(N)}(\omega)[0,H^r) < \infty$. Once again, by Lemma A 7.5 of [12], we infer that for every $\varepsilon > 0$, there exists $l(\omega,\varepsilon) < \infty$ such that $\sup_N \overline{\eta}_0^{(N)}(\omega)(l(\omega,\varepsilon), H^r) < \varepsilon$. Since $\{\langle \mathbf{1}, \overline{\eta}^{(N)}\rangle\}$ is tight by Lemma 6.4, the argument for tightness of $\{\overline{\nu}^{(N)}\}$ presented in Lemma 5.12 of [14] can also be adapted to show that the sequence $\{\overline{\eta}^{(N)}\}$ satisfies Jakubowski's J1 criterion, and is therefore tight. We omit the details. □

We end this section by establishing the relative compactness of the measure-valued processes associated with the cumulative departure and reneging functionals and their compensators.

LEMMA 6.7. *Suppose Assumption 3.1 holds. Then the sequences* $\{\overline{D}^{(N)}_{\cdot}\}$ *and* $\{\overline{A}^{(N)}_{\cdot,\nu}\}$ *are relatively compact in* $\mathcal{D}_{\mathcal{M}_F([0,H^s)\times\mathbb{R}_+)}[0,\infty)$. *Similarly, the sequences* $\{\overline{S}^{(N)}_{\cdot}\}$ *and* $\{\overline{A}^{(N)}_{\cdot,\eta}\}$ *are relatively compact in* $\mathcal{D}_{\mathcal{M}_F([0,H^r)\times\mathbb{R}_+)}[0,\infty)$.

PROOF. This can be proved by combining Lemma 6.3 and Proposition 5.1 with the argument that was used in Lemma 5.13 of [14] to establish the tightness of the sequences $\{\overline{\mathcal{Q}}^{(N)}\}$ and $\{\overline{\mathcal{A}}^{(N)}\}$ therein. Since the adaptation of the argument in [14] is straightforward, we omit the details. □

7. Strong law of large numbers limits.

7.1. *Characterization of subsequential limits.* The focus of this section is the following theorem which, in particular, establishes existence of a solution to the fluid equations.

THEOREM 7.1. *Suppose that Assumptions 3.1–3.3 hold. Let* $(\overline{X},\overline{\nu},\overline{\eta})$ *be the limit of any subsequence of* $\{\overline{X}^{(N)},\overline{\nu}^{(N)},\overline{\eta}^{(N)}\}$. *Then* $(\overline{X},\overline{\nu},\overline{\eta})$ *solves the fluid equations.*

The rest of the section is devoted to the proof of this theorem. Let $(\overline{E},\overline{X}(0),\overline{\nu}_0,\overline{\eta}_0)$ be the $\mathcal{S}_0$-valued random variable that satisfies Assumption 3.1, and let $\{\overline{Y}^{(N)}\}_{N\in\mathbb{N}}$ be the sequence of processes defined in (6.1). Then, by Assumption 3.1, Theorem 6.1 and the limits $\overline{M}^{(N)}_{\cdot,\nu}=\overline{D}^{(N)}_{\cdot}-\overline{A}^{(N)}_{\cdot,\nu}\Rightarrow 0$ and $\overline{M}^{(N)}_{\cdot,\eta}=\overline{S}^{(N)}_{\cdot}-\overline{A}^{(N)}_{\cdot,\eta}\Rightarrow 0$ established in Proposition 5.1, there exist processes $\overline{X}\in\mathcal{D}_{\mathbb{R}_+}[0,\infty)$, $\overline{R}\in\mathcal{D}_{\mathbb{R}_+}[0,\infty)$, $\overline{\nu}\in\mathcal{D}_{\mathcal{M}_F[0,H^s)}[0,\infty)$, $\overline{\eta}\in\mathcal{D}_{\mathcal{M}_F[0,H^r)}[0,\infty)$, $\overline{A}_{\cdot,\nu}\in\mathcal{D}_{\mathcal{M}_F([0,H^s)\times\mathbb{R}_+)}[0,\infty)$, $\overline{D}_{\cdot}\in\mathcal{D}_{\mathcal{M}_F([0,H^s)\times\mathbb{R}_+)}[0,\infty)$, $\overline{A}_{\cdot,\eta}\in\mathcal{D}_{\mathcal{M}_F([0,H^r)\times\mathbb{R}_+)}[0,\infty)$, $\overline{S}_{\cdot}\in\mathcal{D}_{\mathcal{M}_F([0,H^r)\times\mathbb{R}_+)}[0,\infty)$ such that $\overline{Y}^{(N)}$ converges weakly (along a suitable subsequence) to

$$\overline{Y}\doteq(\overline{X}(0),\overline{E},\overline{X},\overline{R},\overline{\nu}_0,\overline{\nu},\overline{\eta}_0,\overline{\eta},\overline{A}_{\cdot,\nu},\overline{A}_{\cdot,\nu},\overline{A}_{\cdot,\eta},\overline{A}_{\cdot,\eta})\in\mathcal{Y}.$$

Denoting this subsequence again by $\overline{Y}^{(N)}$ and invoking the Skorokhod representation theorem, with a slight abuse of notation, we can assume that, $\mathbb{P}$ a.s., $\overline{Y}^{(N)}\to\overline{Y}$ as $N\to\infty$. Without loss of generality, we may further assume that the above convergence holds everywhere.

We now identify some properties of the limit that will be used to prove Theorem 7.1. From Proposition 5.1(1), it follows that, as $N \to \infty$, $(\overline{Y}^{(N)}, \overline{D}^{(N)}_{\cdot}) \to (\overline{Y}, \overline{A}_{\cdot,\nu})$. Together with (2.12), this implies that

$$\overline{X} = \overline{X}(0) + \overline{E} - \overline{A}_{\mathbf{1},\nu} - \overline{R}. \tag{7.1}$$

Moreover, we claim that

$$\overline{A}_{\varphi,\nu} = \int_0^{\cdot} \langle \varphi h^s, \overline{\nu}_s \rangle \, ds. \tag{7.2}$$

This corresponds to relation (5.48) established in Proposition 5.17 of [14] for the model without abandonments. However, essentially the same argument can be used here as well. Specifically, the proof of (5.48) in [14] relies on Lemmas 5.8(1) and 5.16 of [14], which continue to be valid in the presence of abandonments due to Remarks 5.2 and 4.7. On substituting (7.2) into (7.1), we see that the fluid equation (3.15) is satisfied.

Next, in Proposition 7.2, we establish representation (3.14) for $\overline{R}$ given in the fluid equations. The proof of this result relies on the alternative representation for the compensator $A^{(N)}_{\theta^{(N)},\eta}$ of $R^{(N)}$ given in (5.17).

PROPOSITION 7.2. *For every $T \in [0, \infty)$, as $N \to \infty$,*

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\overline{A}^{(N)}_{\theta^{(N)},\eta}(t) - \int_0^t \left(\int_0^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y))\, dy\right) ds\right|\right] \to 0. \tag{7.3}$$

Moreover, almost surely,

$$\overline{R}(t) = \int_0^t \left(\int_0^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y))\, dy\right) ds, \qquad t \in [0, \infty). \tag{7.4}$$

The proof of Proposition 7.2 is given near the end of this section and relies on the following preliminary observations. Let $\tilde{R}(t)$ be defined by the right-hand side of (7.4) for $t \in [0, \infty)$. We first show how (7.4) can be deduced from (7.3). From (7.3), it follows that $\overline{A}^{(N)}_{\theta^{(N)},\eta} \Rightarrow \tilde{R}$ as $N \to \infty$. Since $\tilde{R}$ is continuous, $\overline{R}^{(N)} = \overline{M}^{(N)}_{\theta^{(N)},\eta} + \overline{A}^{(N)}_{\theta^{(N)},\eta}$ and $\overline{M}^{(N)}_{\theta^{(N)},\eta} \Rightarrow 0$ by Lemma 5.4, it follows that $\overline{R}^{(N)} \Rightarrow \tilde{R}$. This implies, a.s., $\tilde{R} = \overline{R}$, and thus the second statement of Proposition 7.2 follows from the first statement.

The proof of (7.3) relies on Lemmas 7.3–7.6 below and the following observations. Using (5.17) and the elementary relation $(F^{\eta_s^{(N)}})^{-1}(N\cdot) = (F^{\overline{\eta}_s^{(N)}})^{-1}(\cdot)$, simple algebraic manipulations show that

$$\overline{A}^{(N)}_{\theta^{(N)},\eta}(t) \doteq \int_0^t \left(\int_0^{\overline{Q}^{(N)}(t)+\overline{\iota}^{(N)}(t)} h^r((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\, dy\right) ds, \qquad t \in [0, \infty), \tag{7.5}$$

where, as usual, $\bar{\iota}^{(N)} \doteq \iota^{(N)}/N$ and $\iota^{(N)}$ is given by (5.15). Next, observe that for all $t \in [0, T]$ and $L \in [0, H^r)$,

$$|\overline{A}^{(N)}_{\theta^{(N)},\eta}(t) - \tilde{R}(t)| \le \overline{C}^{(N)}_1(t, L) + \overline{C}^{(N)}_2(t, L) + \overline{C}_3(t, L), \tag{7.6}$$

where $\overline{C}^{(N)}_i(t, L)$, $i = 1, 2$, and $\overline{C}_3(t, L)$ are defined, for $t \in [0, \infty)$, by

$$\begin{aligned}\overline{C}^{(N)}_1(t, L) \doteq \bigg| &\int_0^t \bigg(\int_0^{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s)) \wedge F^{\overline{\eta}^{(N)}_s}(L)} h^r((F^{\overline{\eta}^{(N)}_s})^{-1}(y))\,dy\bigg)\,ds \\ &- \int_0^t \bigg(\int_0^{\overline{Q}(s) \wedge F^{\overline{\eta}_s}(L)} h^r((F^{\overline{\eta}_s})^{-1}(y))\,dy\bigg)\,ds\bigg|,\end{aligned} \tag{7.7}$$

$$\overline{C}^{(N)}_2(t, L) \doteq \bigg|\int_0^t \bigg(\int_{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s)) \wedge F^{\overline{\eta}^{(N)}_s}(L)}^{\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s)} h^r((F^{\overline{\eta}^{(N)}_s})^{-1}(y))\,dy\bigg)\,ds\bigg| \tag{7.8}$$

and

$$\overline{C}_3(t, L) \doteq \int_0^t \bigg(\int_{\overline{Q}(s) \wedge F^{\overline{\eta}_s}(L)}^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y))\,dy\bigg)\,ds. \tag{7.9}$$

As a precursor to the proof of (7.3) of Proposition 7.2, we first establish some path properties of the limiting queue measure $\overline{\eta}$ in Lemma7.3 and some estimates in Lemma 7.4. These two preliminary results will be used in Lemma 7.5 to show that for any $L \in [0, H^r)$, $\lim_{N\to\infty} \sup_{t\in[0,T]} |\overline{C}^{(N)}_1(t, L)| = 0$ in the case when h^r is continuous. Next, Lemma 7.6 extends this to include general h^r that is locally integrable in $[0, H^r)$. All these results are then combined to prove Proposition 7.2.

LEMMA 7.3. *For every $L \in [0, H^r)$, $\overline{\eta}_t$ is continuous at L for almost every $t \ge 0$. Moreover, for $t \in (0, \infty)$ and $L \in [0, H^r)$, if $\overline{\eta}_t(\{L\}) > 0$, then $\overline{\eta}_t(L, L + \varepsilon) > 0$ for all sufficiently small ε.*

PROOF. It was shown in Corollary 4.2 that $(\overline{\eta}, \overline{E})$ satisfies (3.20) for every bounded Borel measurable function f. For every $L \in [0, H^r)$, substituting $f = \mathbb{1}_L$ in (4.2), we obtain

$$\begin{aligned}\overline{\eta}_t(\{L\}) = &\int_{[0,H^r)} \mathbb{1}_{\{L\}}(x + t)\frac{1 - G^r(x + t)}{1 - G^r(x)}\overline{\eta}_0(dx) \\ &+ \int_{[0,t]} \mathbb{1}_{\{L\}}(t - s)\big(1 - G^r(t - s)\big)\,d\overline{E}(s).\end{aligned} \tag{7.10}$$

It is easy to see that the right-hand side of the above display is zero except when $\overline{\eta}_0(\{L - t\}) > 0$ if $t \le L$ or when $\overline{E}(t - L) - \overline{E}((t - L)-) > 0$ if $t > L$. Since the jump times of both $\overline{\eta}_0$ and $\overline{E}$ are at most countable, (7.10) shows that $\overline{\eta}_t$ is continuous at L for almost every $t \ge 0$.

Next, suppose $\overline{\eta}_t(\{L\}) > 0$. Then by (7.10), at least one of the following two inequalities must hold:

$$\int_{[0,H^r)} \mathbb{1}_{\{L\}}(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\overline{\eta}_0(dx) > 0 \tag{7.11}$$

or

$$\int_{[0,t]} \mathbb{1}_{\{L\}}(t-s)\big(1-G^r(t-s)\big)\,d\overline{E}(s) > 0. \tag{7.12}$$

If (7.11) holds, then it must be that $L-t \in [0, H^r)$, $(1-G^r(L))/(1-G^r(L-t)) > 0$ and $\overline{\eta}_0(\{L-t\}) > 0$. By Assumption 3.2 and the continuity of $(1-G^r(\cdot + t))/(1-G^r(\cdot))$, it then follows that for all sufficient small $\varepsilon > 0$,

$$\int_{[0,H^r)} \mathbb{1}_{(L,L+\varepsilon)}(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\overline{\eta}_0(dx) > 0. \tag{7.13}$$

Substituting $f = \mathbb{1}_{(L,L+\varepsilon)}$ into (3.20) in Corollary 4.2 shows that $\overline{\eta}_t(L, L+\varepsilon)$ is greater than or equal to the left-hand side of (7.13), and so the lemma is established in this case. On the other hand, suppose (7.12) holds. In this case, $t - L > 0$, $1 - G^r(t-L) > 0$ and $\overline{E}(t-L) - \overline{E}((t-L)-) > 0$. By Assumption 3.2 and the continuity of $1 - G^r(t-\cdot)$, for all sufficiently small $\varepsilon > 0$, $1 - G^r(t-\cdot)$ is strictly positive on $(L, L+\varepsilon)$ and $\overline{E}((t-L)-) - \overline{E}(t-L-\varepsilon) > 0$. Another application of (3.20) of Corollary 4.2, with $f = \mathbb{1}_{(L,L+\varepsilon)}$, shows that

$$\overline{\eta}_t(L, L+\varepsilon) \geq \int_0^t \mathbb{1}_{(L,L+\varepsilon)}(t-s)\big(1-G^r(t-s)\big)\,d\overline{E}(s) > 0,$$

and the proof of the lemma is complete. □

LEMMA 7.4. *Let $T \in [0,\infty)$ and $L \in [0, H^r)$. The following estimates hold*:

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

$$\left|\int_0^T \langle \ell, \overline{\eta}_s\rangle\,ds\right| \leq \tilde{L}(m,T)\int_{[0,H^r)} |\ell(x)|\,dx. \tag{7.14}$$

2. *Suppose h is a measurable function such that $\tilde{C}^h_L \doteq \sup_{x\in[0,L]} |h(x)| < \infty$. Then, $\mathbb{P}$-a.s.,*

$$\sup_N \sup_{s\in[0,T]} \int_0^L h(x)\overline{\eta}^{(N)}_s(dx) \leq \tilde{C}^h_L \sup_N\big(\langle \mathbf{1}, \overline{\eta}^{(N)}_0\rangle + \overline{E}^{(N)}(T)\big) < \infty. \tag{7.15}$$

PROOF. It was established in Lemma 5.16 of [14] that inequality (7.14) holds with $\overline{\eta}$ replaced by the fluid age measure $\overline{\nu}$ associated with a many-server queue without abandonments. The proof follows directly from Proposition 4.15 and the estimate (5.46) of [14]. Since the dynamic equations (2.28) and (3.20) for $\eta^{(N)}$

and $\overline{\eta}$, respectively, are exactly analogous to the dynamic equations for $\nu^{(N)}$ and $\overline{\nu}$. Estimate (5.46) of [14] can be shown to hold for $\overline{\eta}$ using the same argument as in [14]. When combined with Proposition 4.15 of [14], this shows that (7.14) holds. Estimate (7.15) follows directly from (2.13) and Assumption 3.1. $\square$

LEMMA 7.5. *For $T \geq 0$ and all but countably many $L \in [0, H^r)$, given any continuous function h on $[0, \infty)$, as $N \to \infty$, for every realization,*

$$\sup_{t\in[0,T]}\left|\int_0^t \left(\int_0^{(\overline{Q}^{(N)}(s)+\overline{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\,dy\right)ds - \int_0^t\left(\int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((F^{\overline{\eta}_s})^{-1}(y))\,dy\right)ds\right| \to 0. \tag{7.16}$$

PROOF. Fix $\omega \in \Omega$. To ease the notation, we shall suppress ω from the notation. From the convergence of $\overline{\eta}^{(N)}$ to $\overline{\eta}$ and $\overline{Q}^{(N)}$ to $\overline{Q}$, it follows that, as $N \to \infty$, $\overline{\eta}_s^{(N)} \stackrel{w}{\to} \overline{\eta}_s$ and $\overline{Q}^{(N)}(s) \to \overline{Q}(s)$ for almost every $s \geq 0$. Also, by Lemma 7.3, $\overline{\eta}_s$ is continuous at L for almost every $s \geq 0$. Let $s \geq 0$ be a time at which $\overline{\eta}_s^{(N)} \stackrel{w}{\to} \overline{\eta}_s$ and $\overline{Q}^{(N)}(s) \to \overline{Q}(s)$ as $N \to \infty$ and $\overline{\eta}_s$ is continuous at L. Then, as $N \to \infty$, $F^{\overline{\eta}_s^{(N)}}(x) \to F^{\overline{\eta}_s}(x)$ for $x = L$ and all but a countable number of $x \in [0, H^r)$. Therefore, by Theorem 13.6.3 of [23], we have $(F^{\overline{\eta}_s^{(N)}})^{-1} \to (F^{\overline{\eta}_s})^{-1}$ on $[0, F^{\overline{\eta}_s}(H^r-))$ in the M_1 topology. For $s \in [0, T]$, we now show that, as $N \to \infty$,

$$\int_0^{(\overline{Q}^{(N)}(s)+\overline{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\,dy \to \int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((F^{\overline{\eta}_s})^{-1}(y))\,dy. \tag{7.17}$$

From the inequality $|\overline{\iota}^{(N)}| \leq 1/N$, we immediately see that

$$\big(\overline{Q}^{(N)}(s) + \overline{\iota}^{(N)}(s)\big) \wedge F^{\overline{\eta}_s^{(N)}}(L) \to \overline{Q}(s) \wedge F^{\overline{\eta}_s}(L) \qquad \text{as } N \to \infty. \tag{7.18}$$

We now consider the following two cases:

Case 1. $\overline{Q}(s) \wedge F^{\overline{\eta}_s}(L) < F^{\overline{\eta}_s}(H^r-)$. In this case, due to (7.18), for all sufficiently large N, $(\overline{Q}^{(N)}(s) + \overline{\iota}^{(N)}(s)) \wedge F^{\overline{\eta}_s^{(N)}}(L) < F^{\overline{\eta}_s}(H^r-)$. For each $n \in \mathbb{N}$, by Theorem 11.5.1 of [23] and the continuity of h, we obtain for each $t < F^{\overline{\eta}_s}(H^r-)$,

$$\lim_{N\to\infty} \sup_{u\in[0,t]}\left|\int_0^u h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\,dy - \int_0^u h((F^{\overline{\eta}_s})^{-1}(y))\,dy\right| = 0.$$

By the case assumption, this implies, in particular, that

$$\lim_{N\to\infty}\left|\int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\,dy - \int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((F^{\overline{\eta}_s})^{-1}(y))\,dy\right| = 0.$$

On the other hand, (7.18) and the continuity of h show that

$$\lim_{N\to\infty}\int_{(\overline{Q}^{(N)}(s)+\overline{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)}^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\,dy=0.$$

Together, the last two assertions imply (7.17).

Case 2. $\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)=F^{\overline{\eta}_s}(H^r-)$. We first claim that in this case

$$\overline{Q}(s)=F^{\overline{\eta}_s}(L)=F^{\overline{\eta}_s}(H^r-). \tag{7.19}$$

Indeed, $F^{\overline{\eta}_s}(L)\le F^{\overline{\eta}_s}(H^r-)$ because $F^{\overline{\eta}_s}$ is nondecreasing and $L<H^r$, while $\overline{Q}(s)\le\overline{\eta}_s[0,H^r)=F^{\overline{\eta}_s}(H^r-)$ by (3.13). On the other hand, the reverse inequalities $\overline{Q}(s)\ge F^{\overline{\eta}_s}(H^r-)$ and $F^{\overline{\eta}_s}(L)\ge F^{\overline{\eta}_s}(H^r-)$ hold by the case assumption, and so the claim follows. Now, define $\overline{L}\doteq(F^{\overline{\eta}_s})^{-1}(F^{\overline{\eta}_s}(H^r-))$. Then $\overline{L}=(F^{\overline{\eta}_s})^{-1}(F^{\overline{\eta}_s}(L))$ by (7.19). Hence, $\overline{L}\le L$ and

$$F^{\overline{\eta}_s}(\overline{L})=F^{\overline{\eta}_s}(L)=F^{\overline{\eta}_s}(H^r-). \tag{7.20}$$

This implies $\overline{\eta}_s(\overline{L},H^r)=0$, and from the second assertion of Lemma 7.3, it follows that

$$\overline{\eta}_s(\{\overline{L}\})=0. \tag{7.21}$$

The change of variables formula and (7.20) then yield

$$\begin{aligned}\int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((F^{\overline{\eta}_s})^{-1}(y))\,dy&=\int_{[0,H^r)} h(x)\overline{\eta}_s(dx)\\&=\int_{[0,\overline{L}]} h(x)\overline{\eta}_s(dx).\end{aligned} \tag{7.22}$$

Also, by Proposition 5.5 and another application of the change of variables formula, we have

$$\begin{aligned}&\int_0^{(\overline{Q}^{(N)}(s)+\overline{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\,dy\\&\qquad=\int_{[0,\chi^{(N)}(s-)]}\mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx).\end{aligned} \tag{7.23}$$

Expanding the term on the right-hand side of (7.23) and using the inequality $\overline{L}\le L$, we obtain

$$\begin{aligned}&\int_{[0,\chi^{(N)}(s-)]}\mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx)\\&\quad=\int_{[0,\overline{L}]}\mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx)\\&\qquad+\int_{(\chi^{(N)}(s-)\wedge\overline{L},\chi^{(N)}(s-)]}\mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx)\\&\qquad-\int_{(\chi^{(N)}(s-)\wedge\overline{L},\overline{L}]}\mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx).\end{aligned} \tag{7.24}$$

By (7.22) and (7.23), the left-hand side and the first term on the right-hand side of (7.24), respectively, equal the left-hand side and right-hand side of (7.17). Therefore, to prove (7.17) it suffices to show that the second and the third terms on the right-hand side of (7.24) converge to zero, as $N \to \infty$. Recall the constant $\tilde{C}_L^h$ defined in Lemma 7.4. Note that $\tilde{C}_L^h < \infty$ since h is continuous. Therefore, the second term on the right-hand side of (7.24) is bounded above by $\tilde{C}_L^h \overline{\eta}_s^{(N)}(\chi^{(N)}(s-) \wedge \overline{L}, \chi^{(N)}(s-)]$. By (7.21), Portmanteau's theorem and (7.20), it follows that

$$\lim_{N\to\infty} \overline{\eta}_s^{(N)}\big(\chi^{(N)}(s-) \wedge \overline{L}, \chi^{(N)}(s-)\big] \leq \lim_{N\to\infty} \overline{\eta}_s^{(N)}(\overline{L}, H^r) = \overline{\eta}[\overline{L}, H^r) = 0.$$

On the other hand, the absolute value of the third term on the right-hand side of (7.24) is bounded above by $\tilde{C}_L^h \overline{\eta}_s^{(N)}(\chi^{(N)}(s-) \wedge \overline{L}, \overline{L}]$. We now argue by contradiction to show that $\liminf_{N\to\infty} \chi^{(N)}(s-) \geq \overline{L}$ and, consequently, that $\overline{\eta}_s^{(N)}(\chi^{(N)}(s-) \wedge \overline{L}, \overline{L}]$ converges to zero as $N \to \infty$. Indeed, suppose this assertion were false. Then there must exist a subsequence $\{N_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \chi^{(N_k)}(s-) = \overline{L} - \delta$ for some $\delta > 0$. Hence, for k large enough, $\chi^{(N_k)}(s-) < \overline{L} - \delta/2$. By Lemma A.2, we have $\chi^{(N_k)}(s-) \geq \chi^{(N_k)}(s)$. Hence $\overline{\eta}_s^{(N_k)}[0, \overline{L} - \delta/2] \geq \overline{Q}^{(N_k)}(s)$ by (2.6). Sending $k \to \infty$ and using the convergence $\overline{\eta}_s^{(N_k)} \Rightarrow \overline{\eta}_s$, the fact that $[0, \overline{L} - \delta/2]$ is closed and Portmanteau's theorem, we obtain $\overline{\eta}_s[0, \overline{L} - \delta/2] \geq \overline{Q}(s)$. This contradicts the definition of $\overline{L}$, and hence completes the proof of (7.17).

Finally, we deduce (7.16) from (7.17) using the bounded convergence theorem, whose application is justified by the bounds (7.22), (7.23) and the estimate (7.15). □

We now generalize Lemma 7.5 to allow for a general locally integrable (not necessarily continuous) function h^r on $[0, H^r)$.

LEMMA 7.6. *Let $L < H^r$, and let $\overline{C}_1^{(N)}(t, L), t \in [0, \infty), N \in \mathbb{N}$ be defined as in (7.7). Then for every $T \in [0, \infty)$, almost surely for $L < H^r$,*

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \overline{C}_1^{(N)}(t, L) = 0. \tag{7.25}$$

PROOF. Fix $L < H^r$. Since h^r lies in $\mathcal{L}^1_{\text{loc}}[0, H^r)$ and is nonnegative, there exists a sequence of nonnegative continuous functions $\{h_n^r\}_{n\geq 1}$ on $[0, H^r)$ such that $\int_0^L |h^r(x) - h_n^r(x)|\, dx \to 0$ as $n \to \infty$ and h_n^r has common compact support in $[0, H^r)$. For each $n \in \mathbb{N}$, (7.25) holds with h_n^r in place of h^r due to Lemma 7.5. Let $l_n^r = |h_n^r - h^r|$ for each $n \geq 1$. Then, in order to prove (7.25), it clearly suffices to show that the following two limits hold: almost everywhere,

$$\lim_{N\to\infty} \sup_N \int_0^T \left(\int_0^{(\overline{Q}^{(N)}(s)+\overline{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} l_n^r((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\, dy\right) ds = 0 \tag{7.26}$$

and

$$\lim_{N\to\infty}\int_0^T\Big(\int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} l_n^r((F^{\overline{\eta}_s})^{-1}(y))\,dy\Big)\,ds=0. \tag{7.27}$$

We first consider (7.26). By Proposition 5.5, applied to $h=l_n^r$, and the same scaling argument that was used to obtain (7.5), for every $N,n\in\mathbb{N}$,

$$\begin{aligned}&\int_0^T\Big(\int_0^{(\overline{Q}^{(N)}(s)+\overline{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} l_n^r((F^{\overline{\eta}_s^{(N)}})^{-1}(y))\,dy\Big)\,ds\\&\qquad=\int_0^T\Big(\int_{[0,\chi^{(N)}(s-)\wedge L]} l_n^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds\le\int_0^T\Big(\int_{[0,L]} l_n^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds.\end{aligned}$$

By (2.2) and the representation of $\eta^{(N)}$ in (2.3), we have

$$\begin{aligned}&\int_0^T\Big(\int_{[0,L]} l_n^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds\\&\qquad\le\frac{1}{N}\sum_{j=-\mathcal{E}_0^{(N)}+1}^{0}\int_0^T l_n^r(w_j^{(N)}(0)+s)\mathbb{1}_{\{w_j^{(N)}(0)+s<L\wedge r_j\}}\,ds\\&\qquad\quad+\frac{1}{N}\sum_{j=1}^{E^{(N)}(T)}\int_{\zeta_j^{(N)}}^T l_n^r(s-\zeta_j^{(N)})\mathbb{1}_{\{s-\zeta_j^{(N)}<L\}}\,ds\\&\qquad\le\sup_N(\langle 1,\overline{\eta}_0^{(N)}\rangle+\overline{E}^{(N)}(T))\int_0^L l_n^r(x)\,dx.\end{aligned}$$

Since $\sup_N(\langle 1,\overline{\eta}_0^{(N)}\rangle+\overline{E}^{(N)}(t))<\infty$ almost surely, due to Assumption 3.1, and h_n^r converges in $\mathcal{L}^1_{\mathrm{loc}}[0,H^r)$ to h^r, we obtain (7.26). On the other hand, observe that, by (7.14) of Lemma 7.4 applied to $l=l_n^r$,

$$\begin{aligned}\int_0^T\Big(\int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} l_n^r((F^{\overline{\eta}_s})^{-1}(y))\,dy\Big)\,ds&\le\int_0^T\Big(\int_{[0,L]} l_n^r(x)\overline{\eta}_s(dx)\Big)\,ds\\&\le\tilde{L}(L,T)\int_0^L l_n^r(x)\,dx.\end{aligned}$$

By the convergence of h_n^r to h^r in $\mathcal{L}^1_{\mathrm{loc}}[0,H^r)$, the last term on the right-hand side of the above display converges to 0, as $n\to\infty$, and (7.27) follows. □

PROOF OF PROPOSITION 7.2. Given the discussion prior to Lemma 7.3 and, in particular, (7.6), to complete the proof of the proposition, it only remains to show that

$$\lim_{L\to H^r}\limsup_{N\to\infty}\mathbb{E}\Big[\sup_{t\in[0,T]}\overline{C}_i^{(N)}(t,L)\Big]=0,\qquad i=1,2, \tag{7.28}$$

and

$$\lim_{L\to H^r} \mathbb{E}[\overline{C}_3(T,L)] = 0. \tag{7.29}$$

For the case $i = 1$ in (7.28), this follows from Lemma 7.6 and the dominated convergence theorem, whose application is justified because, by (7.22), (7.23) and the fact that $\overline{L} \le L$,

$$\mathbb{E}\Big[\sup_{t\in[0,T]} \overline{C}_1^{(N)}(t,L)\Big] \le \mathbb{E}\Big[\int_0^T \Big(\int_{[0,L]} h^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds\Big] + \mathbb{E}\Big[\int_0^T \Big(\int_{[0,L]} h^r(x)\overline{\eta}_s(dx)\Big)\,ds\Big],$$

which is bounded uniformly in N by (7.15) and Assumption 3.1.

Now, by Remark 5.2, an application of Lemma 5.8(1) of [14] (with ν, h^s and H^s, resp., replaced by η, h^r and H^r, resp.), shows that

$$\lim_{L\to H^r} \sup_N \mathbb{E}\Big[\int_0^t \Big(\int_{[L,H^r)} h^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds\Big] = 0. \tag{7.30}$$

On the other hand, the definition of $\overline{C}_2^{(N)}(T,L)$ in (7.8), when combined with Proposition 5.5 and (7.23), shows that

$$\sup_N \mathbb{E}[\overline{C}_2^{(N)}(T,L)] \le \sup_N \mathbb{E}\Big[\int_0^T \Big(\int_{[L,H^r)} h^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds\Big].$$

Taking the limit, as $L \to H^r$, and invoking (7.30), it follows that (7.28) holds for $i = 2$. Finally, to show (7.29), we see that, by the definition of $\overline{C}_3(T,L)$ in (7.9) and the change of variables formula,

$$\begin{aligned}\mathbb{E}[\overline{C}_3(T,L)] &= \mathbb{E}\Big[\int_0^t \Big(\int_{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)}^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y))\,dy\Big)\,ds\Big] \\ &\le \int_0^t \Big(\int_{[L,H^r)} h^r(x)\overline{\eta}_s(dx)\Big)\,ds.\end{aligned}$$

If h^r is bounded, then (7.29) holds by simply applying the bounded convergence theorem on the right-hand side of the equality in the above display. On the other hand, suppose h^r is lower-semicontinuous on (L^r, H^r) for some $L^r < H^r$. Then, by Theorem A.3.12 of [5] and the fact that $\mathbb{P}$ a.s., $\overline{\eta}_s^{(N)} \overset{w}{\to} \overline{\eta}_s$, as $N \to \infty$, for a.e. $s \in [0,T]$, this implies that for any such s and $L > L^r$,

$$\int_0^t \Big(\int_{[L,H^r)} h^r(x)\overline{\eta}_s(dx)\Big)\,ds \le \liminf_{N\to\infty} \int_0^t \Big(\int_{[L,H^r)} h^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds.$$

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

$$\mathbb{E}[\overline{C}_3(T,L)] \le \liminf_{N\to\infty} \mathbb{E}\Big[\int_0^t \Big(\int_{[L,H^r)} h^r(x)\overline{\eta}_s^{(N)}(dx)\Big)\,ds\Big].$$

Taking the limit as $L \to H^r$, an application of (7.30) shows that (7.29) holds. □

We now prove the main limit result.

PROOF OF THEOREM 7.1. Fix $t \in [0,\infty)$ such that $\overline{\nu}_t^{(N)} \stackrel{w}{\to} \overline{\nu}_t$, $\overline{\eta}_t^{(N)} \stackrel{w}{\to} \overline{\eta}_t$, $\overline{E}^{(N)}(t) \to \overline{E}(t)$, $\overline{X}^{(N)}(t) \to \overline{X}(t)$, $\overline{R}^{(N)}(t) \to \overline{R}(t)$, $\overline{A}_{\cdot,\nu}^{(N)}(t) \stackrel{w}{\to} \overline{A}_{\cdot,\nu}(t)$, $\overline{D}_{\cdot}^{(N)}(t) \stackrel{w}{\to} \overline{A}_{\cdot,\nu}(t)$, $\overline{A}_{\cdot,\eta}^{(N)}(t) \stackrel{w}{\to} \overline{A}_{\cdot,\eta}(t)$, $\overline{S}_{\cdot}^{(N)}(t) \stackrel{w}{\to} \overline{A}_{\cdot,\eta}(t)$ as $N \to \infty$. Since $\overline{Y}^{(N)} \to \overline{Y}$ a.s., this occurs for t outside a countable set. By (7.2), this implies that as $N \to \infty$,

$$\overline{D}_{\varphi}^{(N)}(t) \to \overline{A}_{\varphi,\nu}(t) = \int_0^t \langle \varphi(\cdot,s)h^s(\cdot,s), \overline{\nu}_s\rangle\, ds, \qquad \varphi \in \mathcal{C}_b([0,H^s) \times \mathbb{R}_+). \tag{7.31}$$

An analogous argument also implies that, as $N \to \infty$,

$$\overline{S}_{\psi}^{(N)}(t) \to \overline{A}_{\psi,\eta}(t) = \int_0^t \langle \psi(\cdot,s)h^r(\cdot,s), \overline{\eta}_s\rangle\, ds, \qquad \psi \in \mathcal{C}_b([0,H^r) \times \mathbb{R}_+). \tag{7.32}$$

In particular, when $\varphi = \psi = \mathbf{1}$, the above two displays imply that (3.8) holds. Also, we immediately obtain that, as $N \to \infty$, $\langle \mathbf{1}, \overline{\nu}_t^{(N)}\rangle \to \langle \mathbf{1}, \overline{\nu}_t\rangle$ and $\langle \mathbf{1}, \overline{\eta}_t^{(N)}\rangle \to \langle \mathbf{1}, \overline{\eta}_t\rangle$. When combining with (2.15), (2.30), (2.14), (2.20), (2.12), (2.6), (7.4), this implies that all the equations in Definition 3.3 are satisfied at time t except (3.9) and (3.11).

It only remains to show that (3.9) and (3.11) are also satisfied at time t. We shall just prove (3.11). The same argument will also show that (3.9) holds. Dividing (2.28) by N, we have

$$\begin{aligned}\langle \psi(\cdot,t), \overline{\eta}_t^{(N)}\rangle &= \langle \psi(\cdot,0), \overline{\eta}_0^{(N)}\rangle + \int_0^t \langle \psi_x(\cdot,s) + \psi_s(\cdot,s), \overline{\eta}_s^{(N)}\rangle\, ds \\ &\quad - \overline{S}_{\psi}^{(N)}(t) + \int_{[0,t]} \psi(0,s)\, d\overline{E}^{(N)}(s).\end{aligned}$$

Since $\overline{\eta}_0^{(N)} \stackrel{w}{\to} \overline{\eta}_0$ by Assumption 3.1(4), $\overline{\eta}_s^{(N)} \stackrel{w}{\to} \overline{\eta}_s$ for a.e. $s \in [0,t]$, $\overline{\eta}_t^{(N)} \stackrel{w}{\to} \overline{\eta}_t$ by our choice of t and $\psi(\cdot,t)$ and $\psi_x(\cdot,s) + \psi_s(\cdot,s)$, $s \in [0,t]$, are bounded and continuous, as $N \to \infty$, we have

$$\langle \psi(\cdot,t), \overline{\eta}_t^{(N)}\rangle \to \langle \psi(\cdot,t), \overline{\eta}_t\rangle \quad \text{and} \quad \langle \psi(\cdot,0), \overline{\eta}_0^{(N)}\rangle \to \langle \psi(\cdot,0), \overline{\eta}_0\rangle,$$

and, by the bounded convergence theorem,

$$\int_0^t \langle \psi_x(\cdot,s) + \psi_s(\cdot,s), \overline{\eta}_s^{(N)}\rangle\, ds \to \int_0^t \langle \psi_x(\cdot,s) + \psi_s(\cdot,s), \overline{\eta}_s\rangle\, ds.$$

On the other hand, using an integration-by-parts argument, the facts that $\overline{E}^{(N)}(0) = 0$, $\overline{E}^{(N)} \to \overline{E}$, $\overline{E}$ is nondecreasing and $\psi_s(0, \cdot)$ is bounded and continuous on $[0, t]$, along with the bounded convergence theorem, we see that, as $N \to \infty$,

$$\int_{[0,t]} \psi(0, s)\, d\overline{E}^{(N)}(s) \to \int_{[0,t]} \psi(0, s)\, d\overline{E}(s).$$

Combining the last four displays with (7.32), it follows that (3.11) holds. Then it follows that all fluid equations are satisfied for all but countably many t. By right-continuity (with respect to t) of each of the terms in all fluid equations, we conclude that all fluid equations are a.s. satisfied for all $t \in [0, \infty)$. This completes the proof of the desired result that $(\overline{X}, \overline{\nu}, \overline{\eta})$ satisfies the fluid equations. □

7.2. *Proof of Theorem* 3.8. This section is devoted to the proof of Theorem 3.8. Recall $\mathcal{T}_t^{(N)}(s)$ in (2.17) and its fluid scaled version defined in (3.3). Observe that the virtual waiting time defined in (2.18) can be rewritten in terms of the fluid-scaled quantities as

$$W^{(N)}(t) \doteq \inf\{s \geq 0 : \overline{D}^{(N)}(t+s) - \overline{D}^{(N)}(t) + \overline{\mathcal{T}}_t^{(N)}(s) > \overline{Q}^{(N)}(t)\}. \tag{7.33}$$

We first show that for each $t \in [0, \infty)$, $\overline{\mathcal{T}}_t^{(N)} \Rightarrow \overline{\mathcal{T}}_t$ as $N \to \infty$, where $\overline{\mathcal{T}}_t$ is defined in (3.26). Notice that a customer j who arrived into the system before time t and has not reneged by time t must have a potential waiting time $w_j^{(N)}(u) > u - t$ for all $u > t$ sufficiently small. In addition, for that customer to have reneged from the queue (before entering service) in the period $[t, t+s]$, there must exist a time $u \in [t, t+s]$ such that the customer is still in queue (i.e., has not yet entered service) or, equivalently, such that $w_j^{(N)}(u) < \chi^{(N)}(u-)$, the waiting time of the head-of-the-line customer just prior to u, and the customer reneges, so that the slope of her potential waiting time changes from one to zero. Therefore, for each $s \in [0, \infty)$, $\mathcal{T}_t^{(N)}(s)$ can be alternatively expressed as

$$\mathcal{T}_t^{(N)}(s) = \sum_{u \in [t, t+s]} \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(u)} \mathbb{1}_{\{dw_j^{(N)}/dt(u-)>0, dw_j^{(N)}/dt(u+)=0\}} \times \mathbb{1}_{\{u-t<w_j^{(N)}(u) \leq \chi^{(N)}(u-)\}}.$$

Let

$$\mathcal{T}_t^{(N),1}(s) \doteq \sum_{u \in [t, t+s]} \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(u)} \mathbb{1}_{\{dw_j^{(N)}/dt(u-)>0, dw_j^{(N)}/dt(u+)=0\}} \mathbb{1}_{\{w_j^{(N)}(u) \leq \chi^{(N)}(u-)\}}$$

and

$$\mathcal{T}_t^{(N),2}(s) \doteq \sum_{u \in [t, t+s]} \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(u)} \mathbb{1}_{\{dw_j^{(N)}/dt(u-)>0, dw_j^{(N)}/dt(u+)=0\}} \mathbb{1}_{\{w_j^{(N)}(u) \leq u-t\}}.$$

It is easy to see that $\mathcal{T}_t^{(N)}(s) = \mathcal{T}_t^{(N),1}(s) - \mathcal{T}_t^{(N),2}(s)$, $\mathcal{T}_t^{(N),1}(s) = R^{(N)}(t+s) - R^{(N)}(t)$, $\mathcal{T}_t^{(N),2}(s) \le S^{(N)}(t+s) - S^{(N)}(t)$ and $\mathcal{T}_t^{(N),2}(s+\delta) - \mathcal{T}_t^{(N),2}(s) \le S^{(N)}(t+s+\delta) - S^{(N)}(t+s)$. Therefore, an application of Kurtz's criteria in Proposition 6.2 shows that the relative compactness of the fluid scaled versions $\overline{\mathcal{T}}_t^{(N),1}$ and $\overline{\mathcal{T}}_t^{(N),2}$ of $\mathcal{T}_t^{(N),1}$ and $\mathcal{T}_t^{(N),2}$, respectively, follows from that of $\overline{R}^{(N)}$ and $\overline{S}^{(N)}$ established in Lemma 6.3. By a straightforward adaption of the argument used in Proposition 7.2 to show the convergence of $\overline{R}^{(N)}$ to $\overline{R}$, we can conclude that $\overline{\mathcal{T}}_t^{(N)} \Rightarrow \overline{\mathcal{T}}_t$ as $N \to \infty$.

Recall the application of the Skorokhod representation theorem in Theorem 7.1 to assume, without loss of generality, that $\overline{Y}^{(N)}$ converges a.s. to $\overline{Y}$. Here, we can also assume, in addition, that $\overline{\mathcal{T}}_t^{(N)}(s) \to \overline{\mathcal{T}}_t$ a.s., as $N \to \infty$. Since $\overline{Q}$ is continuous at t and, by (7.2), $\overline{A}_{\mathbf{1},\nu} = \int_0^{\cdot} \langle h^s, \overline{\nu}_s\rangle\, ds$ is continuous by the integral representation, and $\overline{\mathcal{T}}_t$ has continuous paths by definition, it follows that, almost surely, $\overline{Q}^{(N)}(t) \to \overline{Q}(t)$ and for each $T \in [0,\infty)$, as $N \to \infty$,

$$\sup_{s\in[0,T]} \left|\overline{D}^{(N)}(t+s) - \overline{A}_{\mathbf{1},\nu}(t+s)\right| \to 0 \quad \text{and} \quad \sup_{s\in[0,T]} \left|\overline{\mathcal{T}}_t^{(N)}(s) - \overline{\mathcal{T}}_t\right| \to 0.$$

From (7.33), it is easy to see that $W^{(N)}(t) \le (\overline{D}^{(N)})^{-1}(\overline{D}^{(N)}(t) + \overline{Q}^{(N)}(t)) - t$ for each N. By the tightness result established in Theorem 6.1, we know that $\overline{D}^{(N)}(t) + \overline{Q}^{(N)}(t)$ is bounded uniformly in N, and due to Lemma 4.10 of [21] and the assumption that $\overline{A}_{\mathbf{1},\nu}$ is uniformly strictly increasing, we also know that $(\overline{D}^{(N)})^{-1} \to (\overline{A}_{\mathbf{1},\nu})^{-1}$ uniformly on compact sets, as $N \to \infty$. Hence, $W^{(N)}(t)$ is bounded uniformly in N. Therefore, there exists a subsequence, $W^{(N_n)}(t)$, $n \in \mathbb{N}$, that converges to a limit in $[0,\infty)$, which we denote by W^*. From (7.33) and the right-continuity of $\overline{D}^{(N)}$, $\overline{Q}^{(N)}$ and $\overline{\mathcal{T}}_t^{(N)}$, we then have $\overline{D}^{(N_n)}(t + \overline{W}^{(N_n)}(t)) - \overline{D}^{(N_n)}(t) + \overline{\mathcal{T}}_t^{(N_n)}(\overline{W}^{(N_n)}(t)) \ge \overline{Q}^{(N_n)}(t)$. Sending $n \to \infty$, we obtain

$$\overline{A}_{\mathbf{1},\nu}(t+W^*) - \overline{A}_{\mathbf{1},\nu}(t) + \overline{\mathcal{T}}_t(W^*) \ge \overline{Q}(t). \tag{7.34}$$

Together with (3.27), this shows that $\overline{W}(t) \le W^*$. Now, suppose that $\overline{W}(t) < W^*$, and fix w such that $\overline{W}(t) < w < W^*$. Since $\overline{A}_{\mathbf{1},\nu}$ is uniformly strictly increasing and $\overline{\mathcal{T}}_t$ is nondecreasing, the inequality $\overline{W}(t) < w$ implies that $\overline{A}_{\mathbf{1},\nu}(t+w) - \overline{A}_{\mathbf{1},\nu}(t) + \overline{\mathcal{T}}_t(w) > \overline{Q}(t)$. Therefore, for sufficiently large N, we have $\overline{D}^{(N)}(t+w) - \overline{D}^{(N)}(t) + \overline{\mathcal{T}}_t^{(N)}(w) > \overline{Q}^{(N)}(t)$ and hence $W^{(N)}(t) \le w$. In turn, this implies that $W^{(N_n)}(t) \le w$ for sufficiently large $n \in \mathbb{N}$. Sending $n \to \infty$ and using the convergence of $W^{(N_n)}(t)$ to W^*, we then obtain $W^* \le w$. This contradicts the choice of w. Hence $\overline{W}(t) = W^*$, and this proves the desired result.

APPENDIX A: EXPLICIT CONSTRUCTION OF THE STATE PROCESSES

In this section, we construct all state processes and auxiliary processes described in Section 2.2 from the initial data $\{\mathcal{E}_0^{(N)}, X^{(N)}(0), w_j^{(N)}(0), a_j^{(N)}(0), j = -\mathcal{E}_0^{(N)} + 1, \dots, 0\}$, $\{\alpha_E^{(N)}(t), t \in [0,\infty)\}$, $\{v_j, j \in \mathbb{Z}\}$ and $\{r_j, j \in \mathbb{Z}\}$.

Fix N and, for simplicity, we omit the dependence on N in notation. Let $E(0) = 0$. The process E on $[0, \infty)$ can be obtained from α_E using the relation (2.1). Let $\ell = 0$, $\tau_0 = 0$, and let $R(\tau_\ell) = D(\tau_\ell) = K(\tau_\ell) = 0$,

$$Q(\tau_\ell) \doteq [X(\tau_\ell) - N]^+, \tag{A.1}$$

and for $j > E(\tau_\ell)$, let $w_j(\tau_\ell) = a_j(\tau_\ell) = 0$. Now, for $t \in [\tau_\ell, \infty)$, define

$$\chi^\ell(t) \doteq \inf\{x > 0 : \eta_{\tau_\ell}[0, x] \geq Q(\tau_\ell)\} + t - \tau_\ell. \tag{A.2}$$

Also, for $j = -\mathcal{E}_0 + 1, \ldots, 0, \ldots, E(\tau_\ell)$ and $t \in [\tau_\ell, \infty)$, let

$$w_j^\ell(t) \doteq (w_j(\tau_\ell) + t - \tau_\ell) \wedge r_j,$$

$$a_j^\ell(t) \doteq \begin{cases} 0, & \text{if } w_j(\tau_\ell) = r_j \text{ or } w_j(\tau_\ell) \leq \chi^\ell(\tau_\ell), \\ (a_j(\tau_\ell) + t - \tau_\ell) \wedge v_j, & \text{if } \chi^\ell(\tau_\ell) < w_j(\tau_\ell) < r_j, \end{cases}$$

$$\eta_t^\ell \doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \delta_{w_j(t)} \mathbb{1}_{\{dw_j/dt(t+)>0\}},$$

$$\nu_t^\ell \doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \delta_{a_j(t)} \mathbb{1}_{\{da_j/dt(t+)>0\}},$$

$$R^\ell(t) \doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \sum_{s\in[0,t]} \mathbb{1}_{\{w_j(s)\leq\chi^l(s-), dw_j/dt(s-)>0, dw_j/dt(s+)=0\}},$$

$$D^\ell(t) \doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \sum_{s\in[0,t]} \mathbb{1}_{\{da_j/dt(s-)>0, da_j/dt(s+)=0\}}.$$

Next, define

$$\tau_{\ell+1} \doteq \inf\{t > 0 : (D^\ell(t) - D(\tau_\ell)) \wedge (R^\ell(t) - R(\tau_\ell)) \wedge (E(t) - E(\tau_\ell)) > 0\}.$$

For $t \in [\tau_\ell, \tau_{\ell+1})$, let $Y(t) = Y^\ell(t)$ for $Y = w_j, a_j, j \in -\mathcal{E}_0 + 1, \ldots, E(\tau_\ell)$, R, D, η, ν and χ and set $Y(t) = Y(\tau_\ell)$ for $Y = X, Q, w_j, a_j, j > E(\tau_\ell)$. Moreover, define

$$X(\tau_{\ell+1}) \doteq X(\tau_\ell) + E(\tau_{\ell+1}) - E(\tau_\ell) - D(\tau_{\ell+1}) + D(\tau_\ell) \\ - R(\tau_{\ell+1}) + R(\tau_\ell),$$

$$\eta_{\tau_{\ell+1}} \doteq \eta^\ell_{\tau_{\ell+1}} + (E(\tau_{\ell+1}) - E(\tau_\ell))\delta_0,$$

and, if $E(\tau_{\ell+1}) > E(\tau_\ell)$, then $E(\tau_{\ell+1}) = E(\tau_\ell) + 1$, and then let $w_j(\tau_{\ell+1}) \doteq 0$ for $j \in \{E(\tau_\ell) + 1, \ldots, E(\tau_{\ell+1})\}$. In this case, $Q(\tau_{\ell+1})$ and $\chi(\tau_{\ell+1})$ can be defined via equations (A.1) and (A.2), but with ℓ replaced by $\ell + 1$, and the procedure can be reiterated. Now, $\max\{\ell : \tau_\ell \leq t\}$ is bounded by $\mathcal{E}_0 + E(t)$, and is therefore a.s. finite. Therefore, $\tau_\ell \to \infty$ as $\ell \to \infty$, and so the above procedure constructs the above processes on $[0, \infty)$. K and S can then be defined, respectively, via equations (2.14) and (2.13).

For each $j \geq -\mathcal{E}_0^{(N)}$, by the construction, we have

$$w_j(t) = \sum_{E(\ell)\geq j} \mathbb{1}_{[\tau_\ell,\tau_\ell+1)}(t)\big(w_j(\tau_\ell) + t - \tau_\ell\big) \wedge r_j$$
$$= \begin{cases} t \wedge r_j, & \text{if } j = -\mathcal{E}_0^{(N)}, \dots, 0, \\ (t - \zeta_j) \wedge r_j, & \text{otherwise,} \end{cases}$$

where $\zeta_j = \inf\{t > 0 : E(t) = j\}$. Hence the process w_j defined above is indeed the potential waiting time process of customer j. It is also not to hard to see that the process a_j defined above is the age process of customer j and satisfies (2.7). We next show that the process χ constructed above satisfies (2.5). It is easy to see that $\chi(0) = \chi^0(0)$ by (A.2) with $t = 0$ and $\ell = 0$. The $\chi(0)$ satisfies (2.5) for $t = 0$. When $t \in [\tau_0, \tau_1)$, $Q(t) = Q(0)$, $\eta_t = \eta_t^0$ and $\chi(t) = \chi^0(t)$. Then we have

$$\chi^0(t) = \inf\{x > 0 : \eta_{\tau_0}[0, x] \geq Q(\tau_0)\} + t - \tau_0 = \inf\{x > 0 : \eta_t[0, x] \geq Q(t)\}.$$

Hence χ satisfies (2.5) on the interval $[\tau_0, \tau_1)$. By the standard induction argument, we can see that χ satisfies (2.5) for all $t \geq 0$.

For each $t \geq 0$, by the construction, we have

$$\eta_t = \sum_{\ell=0}^{\infty} \mathbb{1}_{[\tau_\ell,\tau_\ell+1)}(t) \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \delta_{w_j(t)} \mathbb{1}_{\{dw_j/dt(t+)>0\}}$$
$$= \sum_{\ell=0}^{\infty} \mathbb{1}_{[\tau_\ell,\tau_\ell+1)}(t) \sum_{j=-\mathcal{E}_0+1}^{E(t)} \delta_{w_j(t)} \mathbb{1}_{\{dw_j/dt(t+)>0\}}$$
$$= \sum_{j=-\mathcal{E}_0+1}^{E(t)} \delta_{w_j(t)} \mathbb{1}_{\{dw_j/dt(t+)>0\}}.$$

This shows that the η constructed satisfies (2.3). A similar argument shows that the processes ν, D and R constructed satisfy (2.8), (2.9) and (2.11), respectively. Finally, K and S satisfy (2.14) and (2.13) by construction.

Recall that, for $t \in [0, \infty)$, $\tilde{\mathcal{F}}_t$ is the σ-algebra generated by

$$\big(\mathcal{E}_0, X(0), \alpha_E(s), w_j(s), a_j(s), j \in \{-\mathcal{E}_0 + 1, \dots, 0\} \cup \mathbb{N}, s \in [0, t]\big\}$$

and $\{\mathcal{F}_t\}$ is the associated completed, right-continuous filtration.

LEMMA A.1. *The processes w_j, a_j, $j \geq -\mathcal{E}_0 + 1$ and $E, R, D, \eta, \nu, \chi, X, Q, K, S$ are càdlàg and $\{\mathcal{F}_t\}$-adapted.*

PROOF. The càdlàg property of those processes follows from the construction. Now we show that all the processes are $\{\mathcal{F}_t\}$-adapted. Indeed, it follows immediately from (2.1), (2.3), (2.8), (2.9) and (2.10) that E, η, ν, D and S are $\mathcal{F}_t$-adapted. We next show that χ is $\mathcal{F}_t$-adapted. By equations (2.4) and (2.5) evaluated at time 0, it follows that $\chi(0)$ is a function of $X(0)$ and η_0 and hence $\mathcal{F}_0$-adapted. Now, let $t > 0$. For each $\ell \geq 0$, by the induction argument, $\chi^\ell(t)$ is $\mathcal{F}_t$-adapted, and τ_ℓ is an $\mathcal{F}_t$-stopping time. Since $\chi_t = \chi_t^\ell$ if $t \in [\tau_\ell, \tau_{\ell+1})$, χ is $\mathcal{F}_t$-adapted. Equations (2.11) and (2.12) show that R and X are $\mathcal{F}_t$-adapted, and it follows from (2.4) and (2.14) that Q and K are $\mathcal{F}_t$-adapted. □

The next lemma establishes some basic properties of $\chi(t)$, the waiting time of the head-of-the-line customer at time t, defined in (2.5).

LEMMA A.2. *χ is piecewise linear with downward jumps that occur when the head-of-the-line customer either enters service (due to a departure from service) or reneges from the queue. Hence, $\chi(t-) \geq \chi(t)$ for every $t \in (0, \infty)$. Moreover, for every $t > 0$, there exists $\varepsilon_t(\omega) \in (0, t)$ such that for all $\tilde{t} \in (t - \varepsilon_t(\omega), t)$, $\chi(t-) - \chi(\tilde{t}-) = t - \tilde{t} > 0$.*

PROOF. By the construction, $\chi_t = \chi_t^\ell$ if $t \in [\tau_\ell, \tau_{\ell+1})$. Since χ^ℓ is linear on $[\tau_\ell, \tau_{\ell+1})$, χ is piecewise linear. Also χ can only jump at $\tau_{\ell+1}$, $\ell \geq 0$. Based on the definition of $\tau_{\ell+1}$, it is not hard to see that χ can only have a downward jump at $\tau_{\ell+1}$ when the head-of-the-line customer either enters service $[D^\ell(\tau_{\ell+1}) - D(\tau_\ell) > 0]$ or reneges from the queue $[R^\ell(\tau_{\ell+1}) - R(\tau_\ell) > 0]$. Then we have $\chi(t-) \geq \chi(t)$ for every $t \in (0, \infty)$. The last statement of the lemma follows from the fact that χ is càdlàg and piecewise linear. □

APPENDIX B: STRONG MARKOV PROPERTY

In this section we show that the state descriptor $V^{(N)} = (\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$ is a strong Markov process with respect to the filtration $\{\mathcal{F}_t^{(N)}, t \geq 0\}$ defined in Section 2.2.4. To ease the notation, we shall suppress the superscript (N) from the notation.

Let $\mathcal{M}_D[0, H^s)$ and $\mathcal{M}_D[0, H^r)$ be the subsets of $\mathcal{M}_F[0, H^s)$ and $\mathcal{M}_F[0, H^r)$, respectively, such that each measure in $\mathcal{M}_D[0, H^s)$ and $\mathcal{M}_D[0, H^r)$ takes the form $\sum_{i=1}^k \delta_{x_i}$. Define

$$\mathcal{V} \doteq \left\{ \begin{array}{c} (\alpha, x, \mu, \pi) \in \mathbb{R}_+ \times \mathbb{Z}_+ \times \mathcal{M}_D[0, H^s) \times \mathcal{M}_D[0, H^r): \\ x \leq \langle \mathbf{1}, \mu \rangle + \langle \mathbf{1}, \pi \rangle, \langle \mathbf{1}, \mu \rangle \leq N \end{array} \right\}, \tag{B.1}$$

where $\mathbb{R}_+$ is endowed with the Euclidean topology d, $\mathbb{Z}_+$ is endowed with the discrete topology ρ and $\mathcal{M}_D[0, H^s)$ and $\mathcal{M}_D[0, H^r)$ are endowed with the weak topology, respectively. The space $\mathcal{V}$ is a closed subset of $\mathbb{R}_+ \times \mathbb{Z}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ and is endowed with the usual product topology. Since $\mathbb{R}_+ \times \mathbb{Z}_+ \times$

$\mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ is a Polish space, then the closed subset $\mathcal{V}$ is also a Polish space. Now, denote

$$V(t) \doteq (\alpha_E(t), X(t), \nu_t, \eta_t), \qquad t \geq 0.$$

It is obvious that V is a $\mathcal{V}$-valued process adapted to the filtration $\{\mathcal{F}_t^V, t \geq 0\}$, the natural filtration generated by V.

For each $y, z \in \mathcal{V}$ and $t \geq 0$, let

$$P_t(y, z) = \mathbb{P}\big(V(t) = z | V(0) = y\big). \tag{B.2}$$

For any measurable function ψ defined on $\mathcal{V}$ and $t \geq 0$, define the function $P_t\psi$ on $\mathcal{V}$ as

$$P_t\psi(y) = \mathbb{E}[\psi(V(t)) | V(0) = y], \qquad y \in \mathcal{V}. \tag{B.3}$$

LEMMA B.1. *The state descriptor V is strong Markov with respect to $\{\mathcal{F}_t, t \geq 0\}$, and hence is strong Markov with respect to $\{\mathcal{F}_t^V, t \geq 0\}$. Moreover, $\{P_t, t \geq 0\}$ in* (B.2) *is the Markov semigroup of V.*

PROOF. To establish the strong Markov property, we shall identify V as a, so-called, piecewise deterministic Markov process (cf. [11]). From the explicit pathwise construction of V in Appendix A, it follows that V is a piecewise deterministic process with jump times $\{\tau_1, \tau_2, \ldots\}$. Each jump time is either the arrival time of a new customers or the time of a service completion or the time to the end of a patience time. Note that, due to the nonidling condition, the time of entry into service of a customer must coincide with either the arrival time of that customer or the time of service completion of another customer. Let $\tau_0 = 0$. For each integer $n \geq 0$, let $P_n = V(\tau_n)$. Then $\{(\tau_n, P_n), n \geq 0\}$ forms a marked point process. For each $n \geq 0$, V evolves in a deterministic fashion on $[\tau_n, \tau_{n+1})$. For each $t \geq 0$ and $y \in \mathcal{V}$ with $y = (\alpha, x, \sum_{i=1}^k \delta_{u_i}, \sum_{j=1}^l \delta_{z_i})$ and $k \leq N$, define

$$\phi_t(y) \doteq \left(\alpha + t, x, \sum_{i=1}^k \delta_{u_i+t}, \sum_{j=1}^l \delta_{z_i+t}\right). \tag{$*$}$$

It is easy to see that

$$\phi_{t+s}(y) = \phi_s(\phi_t(y)), \qquad \phi_0(y) = y,$$

and the map $t \mapsto \phi_t(y)$ is continuous in the interval $[0, \infty)$. For each $t \geq 0$, let

$$\langle t \rangle = \max\{n \geq 1 : \tau_n \leq t\}$$

with the convention that $\max \varnothing = 0$. We can see that

$$V(t) = \phi_{t-\tau_{\langle t \rangle}}(V_{\tau_{\langle t \rangle}}). \tag{B.4}$$

The jump dynamics are captured by $\{r_t(y, C), t \geq 0, y \in \mathcal{V}, C \subset \mathcal{V}\}$. For each $t \geq 0, y \in \mathcal{V}, C \subset \mathcal{V}$, $r_t(y, C)$ is the conditional probability that a jump leads to a state in C, given that the jump occurs at time t from state y. Let $y = (\alpha, x, \sum_{i=1}^k \delta_{u_i}, \sum_{j=1}^l \delta_{z_i})$. Recall that there are only three types of jump times for the process V. Given that V jumps at time t from state y, if we know which type the jump time t is, then we know to which state the process V jumps to. For example, suppose that the number k in the expression of y is less than N, then, at state y, there is at least one idle server. If the jump is due to the new arrival, then the process V will jump to state $(0, x+1, \sum_{i=1}^k \delta_{u_i} + \delta_0, \sum_{j=1}^l \delta_{z_i} + \delta_0)$. Let p_1, p_2, p_3, respectively, be the conditional probability that the jump at time t is due to the arrival of a new customer, service completion of a customer in service, the end of patience time for some customer in the system, respectively, given that the jump occurs at time t from state y. Then the probability measure $r_t(y, \cdot)$ can be easily written from y and $p_i, i = 1, 2, 3$.

The jump time dynamics are captured by the survivor functions $\{\overline{H}_{s,y}(t) : 0 \leq s \leq t, y \in \mathcal{V}\}$, where $\overline{H}_{s,y}(t)$ is the conditional probability that the time for the next jump is more than time t given the state being at y at time s, in other words, for $y = (\alpha, x, \sum_{i=1}^k \delta_{u_i}, \sum_{j=1}^l \delta_{z_i})$,

$$
\begin{aligned}
\overline{H}_{s,y}(t) = {} & \frac{1 - F(\alpha + t - s)}{1 - F(\alpha)} \prod_{i=1}^k \frac{1 - G^s(u_i + t - s)}{1 - G^s(u_i)} \\
& \times \prod_{j=1}^l \frac{1 - G^r(z_j + t - s)}{1 - G^r(z_j)}.
\end{aligned}
\tag{$**$}
$$

It is easy to see that $\overline{H}_{s,y}(t)$ satisfies

$$
\overline{H}_{s,y}(u) = \overline{H}_{s,y}(t) \overline{H}_{t, \phi_{t-s}(y)}(u), \qquad s \leq t \leq u.
$$

Then by Theorem 7.3.2 of [11], V is a piecewise deterministic Markov process constructed from $\{(\tau_n, P_n), n \geq 0\}$ using functions ϕ_t for the deterministic part, survivor functions $\overline{H}_{s,y}$ for jump time distributions and transition probabilities r_t for the jumps. Thus it follows from Theorem 7.5.1 of [11] that V is a strong Markov process. The second part of the lemma follows directly from the definition of the $\{P_t, t \geq 0\}$ in (B.2). $\square$

REMARK B.2. For future purposes, we note that the results of this paper including, in particular, the strong Markov property established above, continue to be valid if the state component $\alpha_E^{(N)}$ introduced in Section 2.1 is, instead, defined as follows:

$$
\alpha_E^{(N)}(s) \doteq \begin{cases} s, & \text{if } E^{(N)}(s) = 0, \\ \inf\{u > s : E^{(N)}(u) > E^{(N)}(s)\} - s, & \text{if } E^{(N)}(s) > 0. \end{cases}
$$

Observe that when $E^{(N)}(s) > 0$, $\alpha_E^{(N)}(s)$ represents the time from s until the next arrival, and if $E^{(N)}$ is a renewal process, then $\alpha_E^{(N)}$ is simply the forward recurrence time process. A minor variation of the proof of Lemma B.1 given above shows that the strong Markov property holds in this case as well. First, the definition of $\phi_t(y)$ should be modified by replacing $\alpha + t$ by $\alpha - t$ in $(*)$. With V, r_t and p_1, p_2, p_3 defined as before, in this case, the probability measure $r_t(y, \cdot)$ can be easily determined from y, the distribution of the remaining time from t to the next arrival and p_i, $i = 1, 2, 3$. Note that if $\alpha > 0$ at time t, then $p_1 = 0$. On the other hand, if $\alpha = 0$ at time t, then V jumps at time t due to the arrival of a new customer, and, hence, $p_1 = 1$. Moreover, given that V jumps at time t from state y, if the type of the jump at time t is known, then it is possible to determine the state to which the process V jumps. For example, suppose that the number k in the expression for y is less than N. Then, at state y, there is at least one idle server. If the jump is due to a new arrival, then the state V will jump to the region $\{c \in [0, \infty) : (c, x+1, \sum_{i=1}^k \delta_{u_i} + \delta_0, \sum_{j=1}^l \delta_{z_i} + \delta_0)\}$ according to the distribution of the time to the next arrival (which is determined by the current state α due to the assumption that $\alpha_E^{(N)}$ is Markov with respect to its own filtration). Once again, the jump time dynamics are captured by the survivor functions, with the only difference that now the ratio $(1 - F(\alpha + t - s))/(1 - F(\alpha))$ on the right-hand side of $(**)$ should be replaced by $\mathbb{1}_{\{\alpha \geq t - s\}}$. The rest of the proof then follows as before.

REFERENCES

[1] BACCELLI, F. and HEBUTERNE, G. (1981). On queues with impatient customers. In *Performance '81* (E. Gelenbe, ed.) 159–179. North-Holland, Amsterdam. MR727303

[2] BASSAMBOO, A., HARRISON, J. M. and ZEEVI, A. (2005). Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Syst.* **51** 249–285. MR2189595

[3] BOXMA, O. J. and DE WAAL, P. R. (1994). Multiserver queues with impatient customers. In *Proceedings of ITC* **14** 743–756. Elsevier, Amsterdam.

[4] 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

[5] DUPUIS, P. and ELLIS, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, New York. MR1431744

[6] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York. MR838085

[7] GANS, N., KOOLE, G. and MANDELBAUM, A. (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.

[8] GARNETT, O., MANDELBAUM, A. and REIMAN, M. I. (2002). Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 208–227.

[9] HARRISON, J. M. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* **7** 20–36.

[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] JACOBSEN, M. (2006). *Point Process Theory and Applications*: *Markov Point and Piecewise Deterministic Processes*. Birkhäuser, Boston, MA. MR2189574

[12] KALLENBERG, O. (1975). *Random Measures*. Akademie Verlag, Berlin. MR0431372

[13] KANG, W. N. and RAMANAN, K. (2010). Asymptotic approximations for the stationary distributions of many-server queues. Preprint. Available at http://arxiv.org/abs/1003.3373.

[14] KASPI, H. and RAMANAN, K. (2010). Law of large numbers limits for many-server queues. *Ann*. *Appl*. *Probab*. To appear.

[15] KASPI, H. and RAMANAN, K. (2010). SPDE limits for many-server queues. Preprint.

[16] MANDELBAUM, A., MASSEY, W. A. and REIMAN, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Syst*. **30** 149–201. MR1663767

[17] MANDELBAUM, A. and MOMCILOVIC, P. (2010). Queues with many servers and impatient customers. Preprint.

[18] ZELTYN, S. and MANDELBAUM, A. (2005). Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Syst*. **51** 361–402. MR2189598

[19] MANDELBAUM, A. and ZELTYN, S. (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Oper. Res*. **57** 1189–1205. MR2583510

[20] PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. *Probability and Mathematical Statistics* **3**. Academic Press, New York. MR0226684

[21] 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

[22] WHITT, W. (2006). Fluid models for multiserver queues with abandonments. *Oper. Res*. **54** 37–54. MR2201245

[23] WHITT, W. (2002). *Stochastic-process Limits*: *An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York. MR1876437

[24] ZHANG, J. (2010). Fluid models of many-server queues with abandonment. Preprint.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF MARYLAND
BALTIMORE COUNTY
1000 HILLTOP CIRCLE
BALTIMORE, MARYLAND, 21250
USA
E-MAIL: wkang@umbc.edu

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