# FLUID LIMITS OF MANY-SERVERS QUEUES WITH RENEGING

WEINING KANG AND KAVITA RAMANAN

Abstract. 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 the number of servers goes to infinity, a law of large numbers (or fluid) limit is established for this pair of processes. The limit is 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



*Date*: December 12, 2008.

2000 *Mathematics Subject Classification.* 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.

Partially supported by the National Science Foundation under Grants DMS-0406191, DMS-0405343, CMMI-0728064.

## 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, customers are assumed to abandon from the queue if the time spent waiting in queue reaches the patience time, which is also assumed to be i.i.d., drawn from another 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 [2, 4, 9]. 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 [9]. 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 [19] 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 [2] (see Sections IV and V.2 therein), while 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 [20]. In addition, approximations for performance measures suggested by these limit theorems were used by Garnett et al. [10] and Mandelbaum and Zeltyn [21] 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 [5, 20], thus providing strong motivation for considering many-server systems with general service and abandonment distributions. One previous work that has taken a step towards incorporating more realistic general service distributions is the insightful paper [26], 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 [26]). In this work, we rigorously identify the functional law of large numbers limit, in the limit as the number of servers goes to infinity, for a many-server queueing system with general service and abandonment distributions starting from general initial conditions.

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. This work generalizes the framework of Kaspi and Ramanan [17], in which the corresponding model without abandonments was considered. The presence of two coupled measure-valued processes, rather than just one as in [17], makes the analysis here significantly more involved. In addition, an important step is the identification of an explicit expression for the cumulative reneging process. As in [17], 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 paper also forms the basis of subsequent work, in which we study the long-time behavior of the fluid limit [15] and also establish functional central limit type approximations for many-server queues with abandonment [16].

It is worthwhile to mention that the models discussed above are relevant when the mean demand of customers is known (or can be accurately learnt 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 [11] (see also [3]), 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 state the dynamical equations governing the evolution of the system in Section 2 (the explicit construction of the state process is relegated to Appendix A). 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 a (unique) solution to the fluid equations exists and is obtained as the asymptotic limit of the sequence of scaled state processes. 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 positive integers, $\mathbb{R}$ is set of real numbers and $\mathbb{R}_+$ the set of non-negative real numbers. For $a, b \in \mathbb{R}$, $a \vee b$ denotes the maximum of $a$ and $b$, $a \wedge b$ the minimum of $a$ and $b$ and the short-hand $a^+$ is used for $a \vee 0$. Given $A \subset \mathbb{R}$ and $a \in \mathbb{R}$, $A - a$ equals the set $\{x \in \mathbb{R} : x + a \in A\}$ and $\mathbb{1}_B$ denotes the indicator function of the set $B$ (that is, $\mathbb{1}_B(x) = 1$ 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 $H \leq \infty$, let $\mathcal{L}^1([0, H))$ and $\mathcal{L}^1_{loc}([0, H))$ represent, respectively, the spaces of integrable and locally integrable functions on $[0, H)$, where for $M < \infty$ a locally integrable function $f$ on $[0, H)$ 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 $\mathbf{1}$ and $\mathbf{0}$, respectively. Given any càdlàg, real-valued function $\varphi$ defined on $E$, 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 $\infty$. In addition, the support of a function $\varphi$ 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

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

. The space of Radon measures on a metric space $E$, endowed with the Borel $\sigma$-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 $\mu$ 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, i.e., a sequence of measures $\{\mu_n\}$ in $\mathcal{M}(E)$ is said to converge to $\mu$ in the vague topology (denoted $\mu_n \xrightarrow{v} \mu$) 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, i.e., a sequence $\{\mu_n\}$ in $\mathcal{M}_F(E)$ is said to converge weakly to $\mu$ (denoted $\mu_n \xrightarrow{w} \mu$) 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 $\delta_x$ will be used to denote the measure with unit mass at the point $x$ and, by some abuse of notation, we will use $\mathbf{0}$ 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 $\mu$ on $[0, H)$, we define

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

We will mostly be interested in the case when $E = [0, H)$ and $E = [0, H) \times \mathbb{R}_+$, for some $M \in (0, \infty]$. To distinguish these cases, we will usually use $f$ to denote generic functions on $[0, H)$ and $\varphi$ 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 $(a, b) \subset [0, M)$, 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 $\mathcal{D}_{\mathcal{H}}[0, T]$ (respectively, $\mathcal{D}_{\mathcal{H}}[0, \infty)$) the space of $\mathcal{H}$-valued, càdlàg functions on $[0, T]$ (respectively, $[0, \infty)$), and we endow this space with the usual Skorokhod $J_1$-topology [22]. Then $\mathcal{D}_{\mathcal{H}}[0, T]$ and $\mathcal{D}_{\mathcal{H}}[0, \infty)$ are also Polish spaces (see [22]). 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 $\sigma$-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

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

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 non-decreasing 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 non-idling, First-Come-First-Serve (FCFS) manner, i.e., 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 non-idling condition, that there cannot simultaneously be a positive queue and an idle server, is satisfied. It is assumed that customers are impatient, and renege from the queue as soon as the amount of time spent in the queue reaches their patience times. 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 $i$th 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 $H^r \doteq \sup\{x \in [0,\infty) : G^r(x) < 1\}$ and $H^s \doteq \sup\{x \in [0,\infty) : G^s(x) < 1\}$.

Let $E^{(N)}$ denote the cumulative arrival process, with $E^{(N)}(t)$ representing the total number of customers that arrive into the system in the time interval $[0,t]$. 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

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

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 non-decreasing, 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, for example, 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;

These assumptions are very general, allowing for a large class of arrival processes.

2.2. **State Descriptor and Dynamical Equations.** As mentioned in Section 1.1, our representation of the state of the system 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, we also include the process $X^{(N)}$, which represents the total number of customers in system (including those in service and those in queue), 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 then decoupled from the service dynamics, making it governed purely by the primitive data, $E^{(N)}$ and $G^r$ and more easily analyzable (see Remark 2.2 for further elaboration of this point).

A precise mathematical description of $\eta^{(N)}$ and $\nu^{(N)}$ is 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, we define a filtration $\{\mathcal{F}_t^{(N)}\}$ in the $N$th system, and show that the state processes and auxiliary processes are all adapted to this filtration. It can, in fact, be shown that the state process is Markovian with respect to this filtration, but, since we do not use this fact, we do not provide a proof.

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 $\zeta_j^{(N)} = (E^{(N)})^{-1}(j) \doteq \inf\{t > 0 : E^{(N)}(t) = j\}$, $j \in \mathbb{N}$, then

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

For $j \in -\mathbb{N} \cap \{0\}$, $w_j^{(N)}$ represents the potential waiting time process of the $(j+1)$th customer to enter 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 or, 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 non-negative 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 $\delta_x$ represents the

Dirac mass at $x$. Then the potential queue measure $\eta_t^{(N)}$ can be written in the form

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

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 non-idling 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

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

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 the inverse is 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 $\eta^{(N)}$:

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

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 $j = -\mathcal{E}_0^{(N)}+1, \cdots, 0$, let $a_j^{(N)}(0)$ represent the age of the $(j+1)$th 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, have been in the system longer than the head-of-the-line customer at time $t$, but have not yet departed. the head-of-the-line customer at time $t$. Therefore, a.s., for $j = -\mathcal{E}_0^{(N)}+1, \cdots, 0, \cdots, E^{(N)}(t)$, $t \geq 0$,

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

Note that the condition in the penultimate line 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 non-negative 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

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

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

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

A similar logic shows that the cumulative potential reneging process $S^{(N)}$ admits the representation

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

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

$$R^{(N)}(t) = \sum_{j=-\mathcal{E}_0^{(N)}+1}^{E^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{w_j^{(N)}(s)\le\chi^{(N)}(s-),\ \frac{dw_j^{(N)}}{dt}(s-)>0,\ \frac{dw_j^{(N)}}{dt}(s+)=0\right\}}, \tag{2.11}$$

where the additional restriction $w_j^{(N)}(s) \le \chi^{(N)}(s-)$ 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

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

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

and

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

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 non-idling condition takes the form

$$N - \langle \mathbf{1}, \nu^{(N)} \rangle = [N - X^{(N)}]^+. \tag{2.17}$$

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

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

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 $[t, t+s]$. The virtual waiting time $W^{(N)}(t)$ of a customer at time $t$ is clearly 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 either due to departure from service or due to reneging of customers that arrived prior to time $t$, we can express the virtual waiting time as

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

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

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

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 $\mathbb{R}_+^2 \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$. 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. 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 $\sigma$-algebra generated by

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

and let $\{\mathcal{F}_t^{(N)}\}$ denote the associated right-continuous filtration, completed with respect to $\mathbb{P}$. In Lemma A.1, it is proved that the state process $(\alpha_E^{(N)}, X^{(N)}, \nu^{(N)}, \eta^{(N)})$ and the processes $E^{(N)}$, $Q^{(N)}$, $S^{(N)}$, $R^{(N)}$, $D^{(N)}$ and $K^{(N)}$ are all $\mathcal{F}_t^{(N)}$-adapted.

2.3. **A Succinct Characterization of the Dynamics.** The main result of this section is Theorem 2.1, which provides equations that more succinctly describe 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 $\varphi$ on $[0, H^s) \times \mathbb{R}_+$, consider the sequence of processes $\{D_\varphi^{(N)}\}$ taking values in $\mathbb{R}_+$, given by

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

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 follows that when $\varphi$ is the constant function $\mathbf{1}$, $D_{\mathbf{1}}^{(N)}$ is exactly the cumulative departure process $D^{(N)}$, i.e.,

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

In an exactly analogous fashion, for any measurable function $\psi$ on $[0, H^r) \times \mathbb{R}_+$, consider the sequence of processes $\{S_\psi^{(N)}\}$ taking values in $\mathbb{R}_+$, given by

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

It follows immediately from (2.21) and the right continuity of the filtration $\{\mathcal{F}_t^{(N)}\}$ that $S_\psi^{(N)}$ is $\{\mathcal{F}_t^{(N)}\}$-adapted. It is easy to see that $S_{\mathbf{1}}^{(N)}$ is the cumulative potential reneging process $S^{(N)}$, i.e.,

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

Moreover, it is easy to see that for any $t \in [0,\infty)$ and bounded, measurable $\varphi$,

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

and likewise, for each $t \in [0,\infty)$ and bounded measurable $\psi$,

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

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

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

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

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

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 $\int_{[0,M)} \varphi(x+r,s)\, \nu_s^{(N)}(dx)$, and likewise for $\eta^{(N)}$.

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

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

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

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

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

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

*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 the equations (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 $\nu^{(N)}$ is exactly the same as in the case without abandonment, which was well-studied in [17]. However, in the presence of reneging, there is a significantly more complicated coupling between $\nu^{(N)}$ and $K^{(N)}$, as captured by the equations (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. However, as shown in the sequel, specifically, in Lemma 5.4, (5.48) and Proposition 7.2, we will exploit the representation (2.25) of $R^{(N)}$ in terms of the "known" quantity $S^{(N)}_{\cdot}$ 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 [17], since the definition of $\nu^{(N)}$ in [17] is equivalent to the definition given in (2.8) here since $da_j^{(N)}(t+)/dt = 0$ for all $j > K^{(N)}(t)$ in [17]. 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 $k = 0, \ldots, \lfloor nt \rfloor$,

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

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

$$\sup_{k=0,\ldots,\lfloor nt \rfloor} \left\langle \mathbf{1}, \eta^{(N)}_{\frac{k+1}{n}} \right\rangle \leq E^{(N)}(t+1) + \mathcal{E}_0^{(N)} < \infty. \tag{2.31}$$

Equation (2.29) follows from (2.12), (2.20) and (2.25), while equation (2.30) is the same as (2.17). □

## 3. Main Results

In this section we summarize our main results. First, in Section 3.1, we introduce the fluid-scaled quantities and state some additional 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, which, in particular, show that the fluid equations uniquely characterize the strong law of large numbers or "fluid" limit of the multi-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

$$\overline{\alpha}_E^{(N)}(t) \doteq \alpha_E^{(N)}(t), \qquad \overline{X}^{(N)}(t) \doteq \frac{X^{(N)}(t)}{N}, \tag{3.32}$$

$$\overline{\nu}_t^{(N)}(B) \doteq \frac{\nu_t^{(N)}(B)}{N}, \qquad \overline{\eta}_t^{(N)}(B) \doteq \frac{\eta_t^{(N)}(B)}{N}, \tag{3.33}$$

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

$$\overline{I}^{(N)} \doteq \frac{I^{(N)}}{N} \text{ for } I = E, D, K, Q, R, S, \mathcal{T}_t \tag{3.34}$$

Recall that $\mathcal{I}_{\mathbb{R}_+}[0,\infty)$ is the subset of non-decreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $f(0) = 0$, $H^s = \sup\{x \in [0,\infty) : G^s(x) < 1\}$ and $H^r = \sup\{x \in [0,\infty) : G^r(x) < 1\}$. Define

$$\mathcal{S}_0 \doteq \left\{ \begin{array}{c} (e,x,\nu,\eta) \in \mathcal{I}_{\mathbb{R}_+}[0,\infty) \times \mathbb{R}_+ \times \mathcal{M}_F[0,H^s) \times \mathcal{M}_F[0,H^r) : \\ 1 - \langle \mathbf{1}, \nu \rangle = [1-x]^+ \end{array} \right\}. \tag{3.35}$$

$\mathcal{S}_0$ 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 $\mathbb{E}\left[\overline{E}^{(N)}(t)\right] \to \mathbb{E}\left[\overline{E}(t)\right] < \infty$ for every $t \in [0,\infty)$;*
2. $\overline{X}^{(N)}(0) \to \overline{X}(0)$ *in $\mathbb{R}_+$ $\mathbb{P}$-a.s.;*
3. $\overline{\nu}_0^{(N)} \xrightarrow{w} \overline{\nu}_0$ *in $\mathcal{M}_F[0,H^s)$;*
4. $\overline{\eta}_0^{(N)} \xrightarrow{w} \overline{\eta}_0$ *in $\mathcal{M}_F[0,H^r)$, and $\mathbb{E}\left[\langle 1, \overline{\eta}_0^{(N)} \rangle\right] \to \mathbb{E}[\langle 1, \overline{\eta}_0 \rangle] < \infty$.*

**Remark 3.1.** If the limits in 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 (3.44) imply that, for every $T \in [0,\infty)$,

$$\sup_{t\in[0,T]} \sup_N \mathbb{E}\left[\overline{X}^{(N)}(0) + \overline{E}^{(N)}(t)\right] \leq \mathbb{E}\left[1 + \langle 1, \overline{\eta}_0^{(N)} \rangle + \overline{E}^{(N)}(T)\right] < \infty. \tag{3.36}$$

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

**Assumption 3.2.** *For each $t \geq 0$, 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.*

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

$$h^r(x) \doteq \frac{g^r(x)}{1 - G^r(x)} \text{ for } x \in [0, H^r), \tag{3.37}$$

$$h^s(x) \doteq \frac{g^s(x)}{1 - G^s(x)} \text{ for } x \in [0, H^s). \tag{3.38}$$

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

$$\int_0^t \langle h^r, \overline{\eta}_s\rangle \, ds < \infty, \qquad \int_0^t \langle h^s, \overline{\nu}_s\rangle \, ds < \infty \tag{3.39}$$

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

$$\begin{aligned}\langle \varphi(\cdot,t), \overline{\nu}_t\rangle &= \langle \varphi(\cdot,0), \overline{\nu}_0\rangle + \int_0^t \langle \varphi_s(\cdot,s), \overline{\nu}_s\rangle \, ds + \int_0^t \langle \varphi_x(\cdot,s), \overline{\nu}_s\rangle \, ds \\ &\quad - \int_0^t \langle h^s(\cdot)\varphi(\cdot,s), \overline{\nu}_s\rangle \, ds + \int_0^t \varphi(0,s)\, d\overline{K}(s),\end{aligned} \tag{3.40}$$

where

$$\overline{K}(t) = \langle \mathbf{1}, \overline{\nu}_t\rangle - \langle \mathbf{1}, \overline{\nu}_0\rangle + \int_0^t \langle h^s, \overline{\nu}_s\rangle \, ds; \tag{3.41}$$

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

$$\begin{aligned}\langle \psi(\cdot,t), \overline{\eta}_t\rangle &= \langle \psi(\cdot,0), \overline{\eta}_0\rangle + \int_0^t \langle \psi_s(\cdot,s), \overline{\eta}_s\rangle \, ds + \int_0^t \langle \psi_x(\cdot,s), \overline{\eta}_s\rangle \, ds \\ &\quad - \int_0^t \langle h^r(\cdot)\psi(\cdot,s), \overline{\eta}_s\rangle \, ds + \int_0^t \psi(0,s)\, d\overline{E}(s);\end{aligned} \tag{3.42}$$

$$\overline{Q}(t) = \overline{X}(t) - \langle \mathbf{1}, \overline{\nu}_t\rangle; \tag{3.43}$$

$$\overline{Q}(t) \le \langle \mathbf{1}, \overline{\eta}_t\rangle; \tag{3.44}$$

$$\overline{R}(t) = \int_0^t \left( \int_0^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y)) dy \right) ds, \tag{3.45}$$

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

$$\overline{X}(t) = \overline{X}(0) + \overline{E}(t) - \int_0^t \langle h^s, \overline{\nu}_s\rangle \, ds - \overline{R}(t); \tag{3.46}$$

and

$$1 - \langle \mathbf{1}, \overline{\nu}_t\rangle = [1 - \overline{X}(t)]^+. \tag{3.47}$$

It immediately follows from (3.43) and (3.47) that for each $t \in [0,\infty)$,

$$\overline{Q}(t) = [\overline{X}(t) - 1]^+. \tag{3.48}$$

Also for future use, we observe that (3.41), (3.43) and (3.46), when combined, show that for every $t \in [0,\infty)$

$$\overline{Q}(0) + \overline{E}(t) = \overline{Q}(t) + \overline{K}(t) + \overline{R}(t). \tag{3.49}$$

We now provide an informal, intuitive explanation for the form of the fluid equations. Equations (3.41), (3.43) and (3.46) are simply mass conservation equations,

that are fluid analogs of (2.14), (2.15) and (2.29), respectively, while (3.44) expresses a bound, whose analog clearly holds in the pre-limit, as can be seen from (2.6). The relation (3.47) is simply the fluid analog of the non-idling condition (2.30). Equations (3.40) and (3.42), 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 remark on the fourth term on the right-hand-side of both (3.40) and (3.42), which characterize the fluid departure rate and potential reneging rate as integrals of the corresponding hazard rate with respect to the age and queue measures. Since $\overline{\nu}_s(dx)$ represents the amount of mass (a 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 $[x, x+dx)$, 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 abandonment in [17] (see proposition 5.17). 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

$$\overline{S}(t) \doteq \int_0^t \langle h^r, \overline{\eta}_s\rangle \, ds \qquad \text{for } t \in [0,\infty), \tag{3.50}$$

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 with waiting time $(F^{\overline{\eta}_s})^{-1}(y)$ 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 given as the integral as specified in (3.45).

To close the section, we state the following simple result. For this, we need the following notation: for any $t \in [0,\infty)$,

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

**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 (3.41), (3.45). (3.43) 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 (3.39)–(3.47). Moreover, if $\overline{\nu}$ and $\overline{\eta}$ satisfy*

*(3.39), then $(\overline{X},\overline{\nu},\overline{\eta})$ is a solution to the fluid equations if and only if, for every $f \in \mathcal{C}_b(\mathbb{R}_+)$,*

$$\int_{[0,H^r)} f(x)\,\overline{\eta}_t(dx) = \int_{[0,H^r)} f(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\,\overline{\eta}_0(dx) + \int_0^t f(t-s)(1-G^r(t-s))\,d\overline{E}(s), \tag{3.51}$$

$$\int_{[0,H^s)} f(x)\,\overline{\nu}_t(dx) = \int_{[0,H^s)} f(x+t)\frac{1-G^s(x+t)}{1-G^s(x)}\,\overline{\nu}_0(dx) + \int_0^t f(t-s)(1-G^s(t-s))\,d\overline{K}(s), \tag{3.52}$$

*where*

$$\overline{K}(t) = [\overline{X}(0)-1]^+ - [\overline{X}(t)-1]^+ + \overline{E}(t) - \int_0^t \left( \int_0^{[\overline{X}(s)-1]^+} h^r\left(\left(F^{\overline{\eta}_s}\right)^{-1}(y)\right)\,dy \right) ds \tag{3.53}$$

*and for all $t \in [0,\infty)$, $\overline{X}$ satisfies $[\overline{X}(t)-1]^+ \le \langle \mathbf{1}, \overline{\eta}_t\rangle$, the non-idling condition (3.47) and*

$$\overline{X}(t) = \overline{X}(0) + \overline{E}(t) - \int_0^t \langle h^s, \overline{\nu}_s\rangle\,ds - \int_0^t \left( \int_0^{[\overline{X}(s)-1]^+} h^r\left(\left(F^{\overline{\eta}_s}\right)^{-1}(y)\right)\,dy \right) ds. \tag{3.54}$$

*Moreover, $\overline{K}$ also satisfies*

$$\overline{K}(t) = \int_0^t (\langle \mathbf{1}, \overline{\nu}_{t-s}\rangle - \langle \mathbf{1}, \overline{\nu}_0\rangle) dU^s(s) + \int_0^t \left( \int_{[0,H^s)} \frac{G^s(x+t-s)-G^s(x)}{1-G^s(x)} \overline{\nu}_0(dx) \right) dU^s(s), \tag{3.55}$$

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

Our next result 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. We now state the main result of the paper.

**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.** *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 (3.39)–(3.47) specified in Theorem 3.5. If $\overline{E}$, $\overline{\nu}_0$ and $\overline{\eta}_0$ are absolutely continuous, then $\overline{\nu}_t$, $\overline{\eta}_t$ and $\overline{X}$ are also absolutely continuous for every $t \in [0,\infty)$.*

*Proof.* Since $\overline{E}$ is absolutely continuous, (3.54) allows us to deduce that $\overline{X}$ is absolutely continuous. Therefore, (3.53) shows that $\overline{K}$ is also absolutely continuous. Then the argument used in proving Lemma 5.18 of [17] can be adapted, together with (3.51) and (3.52), to show that $\overline{\nu}_t$, $\overline{\eta}_t$ 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 $\overline{\chi}(u-) = (F^{\overline{\eta}_u})^{-1}(\overline{Q}(u))$. Therefore, by the same logic used to justify the expression (3.45) for $\overline{R}$ in Definition 3.3, it is natural to conjecture that, for each $t \in [0, \infty)$, the fluid limit $\overline{\mathcal{T}}_t^{(N)}$ equals $\overline{\mathcal{T}}_t$, where for $s \in [0, \infty)$,

$$\begin{aligned} \overline{\mathcal{T}}_t(s) &\doteq \int_t^{t+s} \left( \int_{\overline{\eta}_u[0,u-t]}^{\overline{Q}(u)} h^r((F^{\overline{\eta}_u})^{-1}(y)) dy \right) du \\ &= \int_0^s \left( \int_{\overline{\eta}_{t+u}[0,u]}^{\overline{Q}(t+u)} h^r((F^{\overline{\eta}_{t+u}})^{-1}(y)) dy \right) du. \end{aligned} \tag{3.56}$$

Also, define

$$\overline{W}(t) \doteq \inf \left\{ s \geq 0 : \int_t^{t+s} \langle h^s, \overline{\nu}_u \rangle \, du + \overline{\mathcal{T}}_t(s) \geq \overline{Q}(t) \right\}. \tag{3.57}$$

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 $\dot{f}(t) \geq a$ 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 $\overline{\nu}$ (respectively, $\overline{\eta}$) can be written explicitly in terms of the auxiliary fluid process $\overline{K}$ (respectively, $\overline{E}$). In Section 4.2, these representations are used, along with the non-idling 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 [17], 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 (3.40) and (3.42), respectively.

**Proposition 4.1** ([17]). *Let $G$ be the cumulative distribution function of a probability distribution with density $g$ and hazard rate function $h = g/(1-G)$, let $H \doteq \sup\{x \in [0,\infty) : G(x) < 1\}$. 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 $\varphi \in \mathcal{C}_c((-\infty,H)\times\mathbb{R})$ with $supp(\varphi) \subset [0,m]\times[0,T]$. 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*

$$\int_{[0,M)} f(x)\, \overline{\pi}_t(dx) = \int_{[0,M)} f(x+t)\frac{1-G(x+t)}{1-G(x)}\, \overline{\pi}_0(dx) + \int_0^t f(t-s)(1-G(t-s))\, d\overline{Z}(s), \tag{4.3}$$

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

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

The fluid equations (3.39)–(3.42) 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 and Corollary 4.4 of [17] by using a standard approximation argument.

**Corollary 4.2.** *For every bounded Borel measurable function $f$ and $t \in [0,\infty)$, $(\overline{\nu},\overline{K})$ and $(\overline{\eta},\overline{E})$ satisfy*

$$\int_{[0,H^s)} f(x)\, \overline{\nu}_t(dx) = \int_{[0,H^s)} f(x+t)\frac{1-G^s(x+t)}{1-G^s(x)}\, \overline{\nu}_0(dx) + \int_0^t f(t-s)(1-G^s(t-s))\, d\overline{K}(s), \tag{4.5}$$

*and*

$$\int_{[0,H^r)} f(x)\, \overline{\eta}_t(dx) = \int_{[0,H^r)} f(x+t)\frac{1-G^r(x+t)}{1-G^r(x)}\, \overline{\eta}_0(dx) + \int_0^t f(t-s)(1-G^r(t-s))\, d\overline{E}(s). \tag{4.6}$$

*Moreover, $\overline{K}$ satisfies the renewal equation:*

$$\overline{K}(t) = \langle \mathbf{1}, \overline{\nu}_t\rangle - \langle \mathbf{1}, \overline{\nu}_0\rangle + \int_{[0,H^s)} \frac{G^s(x+t)-G^s(x)}{1-G^s(x)}\, \overline{\nu}_0(dx) + \int_0^t g^s(t-s)\overline{K}(s)\, ds. \tag{4.7}$$

**Remark 4.3.** Strictly speaking, in [17] the cumulative distribution was assumed to be absolutely continuous 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 mass at $\infty$. In fact, in the case $G^r$ has a positive mass at $\infty$ 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 (3.43) and (3.45). As an immediate consequence of (3.45), we have the following elementary property.

**Lemma 4.4.** *For any $0 \le a \le b < \infty$, 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 non-decreasing.

**Lemma 4.5.** *The function $\overline{K}$ is non-decreasing.*

*Proof.* Fix $t \in [0, \infty)$ and $0 \le s < t$. If $\overline{X}(t) \ge 1$, then $\langle \mathbf{1}, \overline{\nu}_t \rangle = 1$ by (3.47), and, by (3.41),

$$\overline{K}(t) - \overline{K}(s) = \langle \mathbf{1}, \overline{\nu}_t \rangle - \langle \mathbf{1}, \overline{\nu}_s \rangle + \int_s^t \langle h^s, \overline{\nu}_l \rangle \, dl \ge 0. \tag{4.8}$$

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 (3.43) and (3.47), $\overline{Q}(v) = 0$ 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 (3.49), it then follows that

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

**Case 2:** There exists $v \in (s, t]$ such that $\overline{X}(v) \ge 1$. Define $l \doteq \sup\{v \le t : \overline{X}(v) \ge 1\}$. Then, clearly $l \in (s, t]$ and $\overline{X}(l-) \ge 1$. Now, (3.45) implies that $\overline{R}$ is continuous and hence, by (3.46), $\overline{X}(v) - \overline{X}(v-) \ge 0$ for every $v \in (0, \infty)$. Therefore, $\overline{X}(l) \ge 1 = \langle \mathbf{1}, \overline{\nu}_l \rangle$ and due to the assumption $\overline{X}(t) < 1$, we must have $l < t$. Then (4.8), with $t$ replaced by $l$, shows that $\overline{K}(l) - \overline{K}(s) \ge 0$. On the other hand, since $\overline{X}(v) < 1$ for all $v \in (l, t]$, the argument in Case 1 above shows that $\overline{K}(t) - \overline{K}(l) \ge 0$. Thus, in this case too, we have $\overline{K}(t) - \overline{K}(s) \ge 0$. □

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 (3.51) 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 $\overline{\nu}$. Choose $\delta > 0$ and define

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

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

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

$$\triangle\overline{K}(t) < \delta \text{ if } \langle \mathbf{1}, \overline{\nu}^1_t \rangle = 1. \tag{4.9}$$

To see why this is true, for $t \in [0,\tau]$, suppose $\langle \mathbf{1}, \overline{\nu}^1_t \rangle = 1$. Since $\langle \mathbf{1}, \overline{\nu}^2_t \rangle \le 1$, we have $\langle \mathbf{1}, \triangle\overline{\nu}_t \rangle \le 0$. When combined with (4.7) and the identity $\triangle\overline{\nu}_0 = 0$, this shows that

$$\triangle\overline{K}(t) = \langle \mathbf{1}, \triangle\overline{\nu}_t \rangle + \int_0^t g^s(t-s)\triangle\overline{K}(s)\,ds \le \int_0^t g^s(t-s)\triangle\overline{K}(s)\,ds. \tag{4.10}$$

Using the fact that $\triangle\overline{K}(s) < \delta$ for all $s \in [0,t)$, it is easy to see (for example, as in the proof of Case 2 in Theorem 4.6 of [17]) that this implies $\triangle\overline{K}(t) < \delta G^s(t) \le \delta$, and the claim follows. On the other hand, the right-continuity of $\overline{K}^1$ and $\overline{K}^2$ imply that $\triangle\overline{K}(\tau) \ge \delta$. When combined with (4.9), (3.43) and (3.47), this shows that

$$\langle \mathbf{1}, \overline{\nu}^1_\tau \rangle = \overline{X}^1(\tau) < 1 \text{ and } \overline{Q}^1(\tau) = 0. \tag{4.11}$$

Now, define

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

Then for every $t \in [r,\tau]$, $\overline{Q}^2(t) \ge \overline{Q}^1(t)$. 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 $0 \le \overline{Q}^2(t_n) < \overline{Q}^1(t_n)$ for each $n \in \mathbb{N}$. Since $\overline{Q}^1$ and $\overline{Q}^2$ are càdlàg, this implies that

$$\overline{Q}^2(r-) \le \overline{Q}^1(r-) \tag{4.12}$$

and, due to (3.43) and (3.47), 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 (4.10), this shows that for $n \in \mathbb{N}$,

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

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

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

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 proof of Case 2 in Theorem 4.6 of [17]) that this implies

$$\triangle\overline{K}(r-) < \delta. \tag{4.13}$$

On the other hand, since (3.49) is satisfied with $(\overline{K}, \overline{R}, \overline{Q})$ replaced by $(\overline{K}^i, \overline{R}^i, \overline{Q}^i)$ for $i = 1, 2$, 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,

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

Since $\triangle\overline{Q}(r-) \le 0$ by (4.12), and $-\triangle\overline{Q}(\tau) = \overline{Q}^1(\tau) - \overline{Q}^2(\tau) = -\overline{Q}^2(\tau) \le 0$ due to (4.11), we obtain

$$\triangle\overline{K}(\tau) - \triangle\overline{K}(r-) \le -(\triangle\overline{R}(\tau) - \triangle\overline{R}(r-)). \tag{4.14}$$

For each $t \geq 0$, by (3.45), we see that

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

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

$$\begin{aligned}(4.15)\quad &\triangle\overline{R}(\tau) - \triangle\overline{R}(r-) \\ &= \triangle\overline{R}(\tau) - \triangle\overline{R}(r) \\ &= \int_r^\tau \left( \int_0^{\overline{Q}^2(s)} h^r((\overline{F}^{\overline{\eta}^1_s})^{-1}(y))dy \right) ds - \int_r^\tau \left( \int_0^{\overline{Q}^1(s)} h^r((\overline{F}^{\overline{\eta}^1_s})^{-1}(y))dy \right) ds.\end{aligned}$$

However, by the definition of $\tau$, 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 (4.14) and (4.13), this implies

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

A similar 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 $\tau$. 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 (4.5) of Corollary 4.2, it follows that $\overline{\nu}^1 = \overline{\nu}^2$. Also, by (3.49), we obtain

$$\overline{R}^1 + \overline{Q}^1 = \overline{R}^2 + \overline{Q}^2. \tag{4.16}$$

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

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

It follows that $\overline{Q}^1(s-) \leq \overline{Q}^2(s-)$ 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

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

It then follows from (4.16) and the continuity of $\overline{R}^i$, $i = 1, 2$, that $\overline{Q}^1(t) - \overline{Q}^1(s-) \leq \overline{Q}^2(t) - \overline{Q}^2(s-)$. Combining this with the inequality $\overline{Q}^1(s-) \leq \overline{Q}^2(s-)$, we obtain $\overline{Q}^1(t) \leq \overline{Q}^2(t)$, which is a contradiction. Hence $\overline{Q}^1(v) \leq \overline{Q}^2(v)$ for all $v \in (0, \infty)$. Similarly 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 $\overline{R}^1 = \overline{R}^2$. In the end, by (3.43), 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 (3.43), (3.45) and (3.46), while the alternative representation for $\overline{K}$ is a direct consequence of the renewal equation (4.7) and Corollary 4.4 of [17]. □

## 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 the model without abandonment, it was shown in [17] that the process $A_{\varphi,\nu}^{(N)}$ defined below is the compensator for the associated "$\varphi$-weighted" cumulative departure process. Since the service dynamics is similar in the model with or without abandonment, properties of the departure process analogous to those established in Section 5.2 of [17] continue to hold in the presence of abandonment. These properties are summarized in Proposition 5.1. Next, we exploit the fact that the relation between the potential reneging process $S^{(N)}$ and the queue measure $\eta^{(N)}$ is exactly analogous to the relation between the departure process $D^{(N)}$ and the age measure $\nu^{(N)}$, in order to identify the compensator of the "$\psi$-weighted" potential reneging process $S_\psi^{(N)}$ for a suitable $\psi$ (see Lemma 5.2). Using Lemma 5.2 and the representation $R^{(N)} = S_{\theta^{(N)}}^{(N)}$ for the reneging process established in (2.25), we then deduce the form of the compensator for $R^{(N)}$ (see Lemma 5.4). Since $\theta^{(N)}$ does not belong to the class of $\psi$ specified in Lemma 5.2, this requires some additional justification which is provided in Lemma 5.3. Finally, by essentially the same arguments used to prove properties of $D^{(N)}$ in Proposition 5.1 (2) and (3), we obtain the corresponding properties for $R^{(N)}$. These are summarized in Proposition 5.5.

**Proposition 5.1.** *Let $D_\varphi^{(N)}$ be the process defined in (2.19). Then the following properties hold.*

(1) *For every bounded measurable function $\varphi$ on $[0, H^s)\times\mathbb{R}_+$ such that $\varphi(a_j^{(N)}(\cdot),\cdot)$ is left continuous for each $j$, the process $A_{\varphi,\nu}^{(N)}$ is the $\mathcal{F}_t^{(N)}$-compensator of the process $D_\varphi^{(N)}$. In particular, the process $M_{\varphi,\nu}^{(N)}$ defined by*

$$M_{\varphi,\nu}^{(N)} \doteq D_\varphi^{(N)} - A_{\varphi,\nu}^{(N)} \tag{5.17}$$

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

$$|A_{\varphi,\nu}^{(N)}(t)| \le \|\varphi\|_\infty \left(X^{(N)}(0) + E^{(N)}(t)\right)\left(\int_0^m h^s(x)\,dx\right) < \infty \tag{5.18}$$

*for every $\varphi \in \mathcal{C}_c([0, H^s)\times\mathbb{R}_+)$ with $\mathrm{supp}(\varphi) \subset [0,m]\times\mathbb{R}_+$. 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*

$$\lim_{N\to\infty} \mathbb{E}\left[\langle\overline{M}_{\varphi,\nu}^{(N)}\rangle(t)\right] = 0. \tag{5.19}$$

*Consequently, as* $N \to \infty$,

$$\overline{M}^{(N)}_{\varphi,\nu} \Rightarrow \mathbf{0}. \tag{5.20}$$

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

$$\limsup_N \mathbb{E}\left[\left|\overline{D}^{(N)}(T)\right|\right] < \infty \text{ and } \limsup_N \mathbb{E}\left[\left|A^{(N)}_{\varphi,\nu}(T)\right|\right] < \infty. \tag{5.21}$$

*Also, for* $t \in [0, \infty)$ *and* $N \in \mathbb{N}$,

$$\lim_{\delta \to 0} \mathbb{E}\left[\overline{D}^{(N)}(t+\delta) - \overline{D}^{(N)}(t)\right] = 0. \tag{5.22}$$

*Moreover, for every* $\delta > 0$ *and interval* $\mathcal{Z} = [L+\delta, H^s)$ *with* $L \in (0, H^s - \delta)$,

$$\mathbb{E}\left[\overline{D}^{(N)}_{\mathbb{1}_{\mathcal{Z}}}(t+\delta) - \overline{D}^{(N)}_{\mathbb{1}_{\mathcal{Z}}}(t) | \mathcal{F}^{(N)}_t\right] \le U^s(\delta)\overline{\nu}^{(N)}_t[L, H^s), \tag{5.23}$$

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

(3) *Suppose that the limit*

$$\lim_{L \to H^s} \sup_{N \in \mathbb{N}} \mathbb{E}\left[\overline{\nu}^{(N)}_0(L, H^s)\right] = 0 \tag{5.24}$$

*holds and, if* $H^s < \infty$, *then*

$$\lim_{L \to H^s} \sup_{N \in \mathbb{N}} \mathbb{E}\left[\int_{[0,L)} \frac{1 - G^s(L)}{1 - G^s(x)} \, \overline{\nu}^{(N)}_0(dx)\right] = 0 \tag{5.25}$$

*is also satisfied. Then the following three properties hold.*

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

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

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

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

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

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

(c) *Given* $L < H^s$ *and any sequence of measurable subsets* $B_R \subset [0, L]$ *such that the Lebesgue measure of* $B_R$ *goes to zero as* $R \to \infty$, *we have for every* $T \in [0, \infty)$,

$$\lim_{R \to \infty} \limsup_N \mathbb{E}\left[\sup_{t \in [0,T]} \overline{A}^{(N)}_{\mathbb{1}_{B_R},\nu}(t)\right] = 0. \tag{5.26}$$

*Proof.* The fact that $A^{(N)}_{\varphi,\nu}$ is the $\mathcal{F}^{(N)}_t$-compensator of $D^{(N)}_{\varphi}$ essentially follows from Lemma 5.4 and Corollary 5.5 in [17]. The only difference is that the filtration $\{\mathcal{F}^{(N)}_t\}$ here is larger than that in [17] since here, $\mathcal{F}^{(N)}_t$ also includes the $\sigma$-algebra generated by the potential waiting times $\{\eta^{(N)}_j(s),\ s \le t,\ j = -\mathcal{E}^{(N)}_0 + 1, \cdots, E^{(N)}(t)\}$. However, due to the assumed independence of the patience times and the service times, the proof of Lemma 5.4 of [17] continues to be

valid in the setting of this paper. In fact, the only argument in Lemma 5.4 of [17] that needs to be checked in the setting of this paper is the following:

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\mathbb{1}_{\{\zeta_n^k\le\frac{j+1}{2^m}\}}|\mathcal{F}_{\frac{j}{2^m}}\right]\\&\quad=\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\int_{j/2^m}^{(j+1)/2^m}\frac{g^s(u-\theta_n^k)}{1-G^s(\frac{j}{2^m}-\theta_n^k)}du,\end{aligned}\tag{5.27}$$

where $\theta_n^k$ (respectively, $\zeta_n^k$) is the time at which the $n$-th customer to be served at station $k$ starts (respectively, completes) service. Then $\zeta_n^k-\theta_n^k$ is the service time of the $n$-th customer to be served at station $k$, which has distribution $G^s$. Let $\mathcal{G}_{\frac{j}{2^m}}$ be the $\sigma$-algebra generated by the events $\{(\theta_n^k\le x)\cap(\theta_n^k\le\frac{j}{2^m},\zeta_n^k>\frac{j}{2^m}),\ x\ge 0\}$. In order to show the equality in (5.27), it suffices to show that for every bounded $\mathcal{F}_{\frac{j}{2^m}}$-adapted random variable $H$,

$$\begin{aligned}&\mathbb{E}\left[H\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\mathbb{1}_{\{\zeta_n^k\le\frac{j+1}{2^m}\}}\right]\\&\quad=\mathbb{E}\left[H\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\int_{j/2^m}^{(j+1)/2^m}\frac{g^s(u-\theta_n^k)}{1-G^s(\frac{j}{2^m}-\theta_n^k)}du\right].\end{aligned}\tag{5.28}$$

Recall that the patience times and the service times of customers are assumed to be independent. Therefore, given $\mathcal{G}_{\frac{j}{2^m}}$, $\zeta_n^k-\theta_n^k$ and $\mathcal{F}_{\frac{j}{2^m}}$ are conditionally independent. Hence, it follows from the left-hand-side of (5.28) and the fact that $H$ is $\mathcal{G}_{\frac{j}{2^m}}$-adapted that

$$\begin{aligned}&\mathbb{E}\left[H\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\mathbb{1}_{\{\zeta_n^k\le\frac{j+1}{2^m}\}}\right]\\&\quad=\mathbb{E}\left[\mathbb{E}\left[H\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\mathbb{1}_{\{\zeta_n^k-\theta_n^k\le\frac{j+1}{2^m}-\theta_n^k\}}|\mathcal{G}_{\frac{j}{2^m}}\right]\right]\\&\quad=\mathbb{E}\left[\mathbb{E}\left[H|\mathcal{G}_{\frac{j}{2^m}}\right]\mathbb{E}\left[\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\mathbb{1}_{\{\zeta_n^k-\theta_n^k\le\frac{j+1}{2^m}-\theta_n^k\}}|\mathcal{G}_{\frac{j}{2^m}}\right]\right]\\&\quad=\mathbb{E}\left[H\mathbb{E}\left[\mathbb{1}_{\{\theta_n^k\le\frac{j}{2^m}<T,\zeta_n^k>\frac{j}{2^m}\}}\mathbb{1}_{\{\zeta_n^k-\theta_n^k\le\frac{j+1}{2^m}-\theta_n^k\}}|\mathcal{G}_{\frac{j}{2^m}}\right]\right],\end{aligned}$$

and

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

This shows that (5.28), and therefore (5.27), holds.

Corollary 5.5 of [17] shows that $M_{\varphi,\nu}^{(N)}$ is a local $\mathcal{F}_t^{(N)}$-martingale for bounded and continuous functions $\varphi$. However, it is easy to see that this holds more generally for bounded and measurable $\varphi$ such that $\varphi(a_j^{(N)}(\cdot),\cdot)$ is left continuous for each $j$, since then $\varphi(a_j^{(N)}(\cdot),\cdot)$ is $\mathcal{F}_t^{(N)}$-predictable. Property 2 can be established exactly as Lemma 5.6 of [17], while property 3 follows from Lemma 5.8 of [17]. □

Suppositions (5.24) and (5.25) above are shown in Lemma 6.6 to follow from Assumption 3.1(3).

We now turn to the cumulative reneging processes. For any bounded measurable function $\psi$ on $[0,H^r)\times\mathbb{R}_+$, consider the sequence $\{A_{\psi,\eta}^{(N)}\}$ of processes given by

$$A_{\psi,\eta}^{(N)}(t)\doteq\int_0^t\left(\int_{[0,H^r)}\psi(x,s)h^r(x)\,\eta_s^{(N)}(dx)\right)ds,\qquad t\in[0,\infty).\tag{5.29}$$

Due to the analogy between the service dynamics and the potential queue dynamics (see Remark 2.2), the i.i.d nature of the sequences of service requirements and patience times, and the independence of these two sequences from each other and the cumulative arrival processes, the argument used in the proof of Proposition 5.1(1) can also be used to show that $A^{(N)}_{\psi,\eta}$ is well-defined and equals the compensator of $S^{(N)}_{\psi}$. Recall by (2.22) that $S^{(N)} = S^{(N)}_{\mathbf{1}}$.

**Lemma 5.2.** *For every $N \in \mathbb{N}$, the process $A^{(N)}_{\mathbf{1},\eta}$ is the $\mathcal{F}^{(N)}_t$-compensator of the process $S^{(N)}$. Hence, for every bounded measurable function $\psi$ on $[0, H^r) \times \mathbb{R}_+$ such that $\psi(w^{(N)}_j(\cdot), \cdot)$ is left continuous for each $j$, the process $A^{(N)}_{\psi,\eta}$ is the $\mathcal{F}^{(N)}_t$-compensator of the process $S^{(N)}_{\psi}$. In particular, for such $\psi$, the process $M^{(N)}_{\psi,\eta}$ defined by*

$$M^{(N)}_{\psi,\eta} \doteq S^{(N)}_{\psi} - A^{(N)}_{\psi,\eta} \tag{5.30}$$

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

$$|A^{(N)}_{\psi,\eta}(t)| \le \|\psi\|_\infty \left(\langle \mathbf{1}, \eta^{(N)}_0 \rangle + E^{(N)}(t)\right) \left(\int_0^m h^r(x)\, dx\right) < \infty \tag{5.31}$$

*for every $\psi \in \mathcal{C}_c([0, m) \times \mathbb{R}_+)$ with $\mathrm{supp}(\psi) \subset [0, m] \times \mathbb{R}_+$ for $m \in [0, H^r)$.*

Now, note from (2.25) that $R^{(N)} = S^{(N)}_{\theta^{(N)}}$, where $\theta^{(N)}$ is defined by (2.26). Also note from (2.25) that $R^{(N)} = S^{(N)}_{\theta^{(N)}}$. Therefore, in view of (5.17), it is natural to conjecture that the compensator of $R^{(N)}$ is equal to $A^{(N)}_{\theta^{(N)},\eta}$, where

$$A^{(N)}_{\theta^{(N)},\eta}(t) \doteq \int_0^t \left( \int_{[0,H^r)} \mathbb{1}_{[0,\chi^{(N)}(s-)]}(x) h^r(x)\, \eta^{(N)}_s(dx) \right) ds, \qquad t \in [0, \infty). \tag{5.32}$$

However, this is not immediate from Lemma 5.2 since $\theta^{(N)}(w^{(N)}_j(\cdot), \cdot)$ is not left continuous for any $j$. Instead, we approximate $\theta^{(N)}$ by a sequence $\{\theta^{(N)}_m\}_{N\in\mathbb{N}}$ defined by

$$\theta^{(N)}_m(x, s) \doteq \mathbb{1}_{(x-\frac{1}{m},\infty)}(\chi^{(N)}(s-)), \tag{5.33}$$

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 $m \ge 1$, $x \in \mathbb{R}$ and $s \in \mathbb{R}_+$, the sequence $\{\theta^{(N)}_m\}_{N\in\mathbb{N}}$ defined by (5.33) satisfies the following two properties:*

1. *For every $N \in \mathbb{N}$, $x \in \mathbb{R}$, $s \in \mathbb{R}$, $\theta^{(N)}_m(x, s)$ is non-increasing in $m$ and converges, as $m \to \infty$, to $\theta^{(N)}(x, s)$.*
2. *For each $N, m \in \mathbb{R}$, $j \in \mathbb{Z}$, the process $\theta^{(N)}_m(w^{(N)}_j(\cdot), \cdot)$ has left continuous paths on $(0, \infty)$.*

*Proof.* The first property is immediate from the definition of $\theta^{(N)}_m$. 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 $\omega$ 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 non-decreasing, $w_j^{(N)}(s_n) \le w_j^{(N)}(s)$ and since $\chi^{(N)}(s-)$ is left continuous, we have, for all $n$ large enough, $w_j^{(N)}(s_n) < \chi^{(N)}(s_n-)+1/m$. Hence, $\theta_m^{(N)}(w_j^{(N)}(s_n), s_n) = 1$ and $\theta_m^{(N)}(w_j^{(N)}(\cdot), \cdot)$ 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 $w_j^{(N)}(s) - w_j^{(N)}(s_n) \le s - s_n$ 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 $s$. □

**Lemma 5.4.** *For every $N \in \mathbb{N}$, the process $A^{(N)}_{\theta^{(N)},\eta}$ is the $\mathcal{F}_t^{(N)}$-compensator of the process $R^{(N)}$. In particular, the process $M^{(N)}_{\theta^{(N)},\eta}$ defined by*

$$M^{(N)}_{\theta^{(N)},\eta} \doteq R^{(N)} - A^{(N)}_{\theta^{(N)},\eta} \tag{5.34}$$

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

*Proof.* Fix $N \in \mathbb{N}$, and let $A^{(N)}_{\theta_m^{(N)},\eta}$, $m \in \mathbb{N}$, be defined in the obvious way:

$$A^{(N)}_{\theta_m^{(N)},\eta}(t) \doteq \int_0^t \left( \int_{[0,H^r)} \theta_m^{(N)}(x,s) h^r(x)\, \eta_s^{(N)}(dx) \right) ds. \tag{5.35}$$

By Lemma 5.2 and Lemma 5.3, the process $A^{(N)}_{\theta_m^{(N)},\eta}$ is the $\mathcal{F}_t^{(N)}$-compensator of the process $S^{(N)}_{\theta_m^{(N)}}$, and the process $M^{(N)}_{\theta_m^{(N)},\eta}$ defined by

$$M^{(N)}_{\theta_m^{(N)},\eta} \doteq S^{(N)}_{\theta_m^{(N)}} - A^{(N)}_{\theta_m^{(N)},\eta} \tag{5.36}$$

is a local $\mathcal{F}_t^{(N)}$-martingale. Since $\theta_m^{(N)} \to \theta^{(N)}$ 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 $\mathbb{E}\left[S^{(N)}_{\mathbf{1}}(t)\right] < \infty$, $\mathbb{E}\left[A^{(N)}_{\mathbf{1},\eta}(t)\right] < \infty$ for all $t \in (0,\infty)$, an application of the dominated convergence theorem shows that for all $t \in (0,\infty)$, as $m \to \infty$,

$$\mathbb{E}\left[\sup_{0\le s\le t} \left| A^{(N)}_{\theta_m^{(N)},\eta}(s) - A^{(N)}_{\theta^{(N)},\eta}(s) \right| \right] \to 0 \text{ and } \mathbb{E}\left[\sup_{0\le s\le t} \left| S^{(N)}_{\theta_m^{(N)}}(s) - S^{(N)}_{\theta^{(N)}}(s) \right| \right] \to 0,$$

and hence $M^{(N)}_{\theta_m^{(N)},\eta}$ converges in law to $M^{(N)}_{\theta^{(N)},\eta}$. Since $\left| S^{(N)}_{\theta_m^{(N)}}(t) - S^{(N)}_{\theta_m^{(N)}}(t-) \right| \le 1$ for all $t \in [0,\infty)$ and $m \in \mathbb{N}$, then $M^{(N)}_{\theta^{(N)},\eta}$ is a local $\mathcal{F}_t^{(N)}$-martingale by Corollary 1.19 of Chapter IX of [12]. □

As usual, let $\overline{A}^{(N)}_{\psi,\eta}$, $\overline{M}^{(N)}_{\psi,\eta}$ and $\overline{A}^{(N)}_{\theta^{(N)},\eta}$, $\overline{M}^{(N)}_{\theta^{(N)},\eta}$ denote the scaled versions $A^{(N)}_{\psi,\eta}/N$, $M^{(N)}_{\psi,\eta}/N$ and $A^{(N)}_{\theta^{(N)},\eta}/N$, $M^{(N)}_{\theta^{(N)},\eta}/N$, respectively. In the next proposition, we collect some useful properties of these processes. Recall that the scaled reneging process $\overline{R}^{(N)}$ equals $\overline{A}^{(N)}_{\theta^{(N)},\eta}$.

**Proposition 5.5.** *The following properties hold.*

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

$$\lim_{N\to\infty}\mathbb{E}\left[\langle \overline{M}^{(N)}_{\psi,\eta}\rangle(t)\right]=0. \tag{5.37}$$

*Moreover,*

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

*Consequently,* $\overline{M}^{(N)}_{\psi,\eta}\Rightarrow \mathbf{0}$ *and* $\overline{M}^{(N)}_{\theta^{(N)},\eta}\Rightarrow \mathbf{0}$ *as* $N\to\infty$.

(2) *For every* $T<\infty$ *and* $\psi\in\mathcal{C}_b([0,H^r)\times\mathbb{R}_+)$,

$$\sup_N \mathbb{E}\left[\overline{R}^{(N)}(T)\right]<\infty \text{ and } \limsup_N \mathbb{E}\left[\left|A^{(N)}_{\psi,\eta}(T)\right|\right]<\infty. \tag{5.39}$$

*Also, for* $t\in[0,\infty)$ *and* $N\in\mathbb{N}$,

$$\lim_{\delta\to 0}\mathbb{E}\left[\overline{R}^{(N)}(t+\delta)-\overline{R}^{(N)}(t)\right]=0. \tag{5.40}$$

*For every* $\delta>0$ *and interval* $\mathcal{Z}=[L+\delta,H^r)$ *with* $L\in(0,H^r-\delta)$,

$$\mathbb{E}\left[\overline{S}^{(N)}_{\mathbb{1}_{\mathcal{Z}}}(t+\delta)-\overline{S}^{(N)}_{\mathbb{1}_{\mathcal{Z}}}(t)|\mathcal{F}^{(N)}_t\right]\le U^r(\delta)\overline{\eta}^{(N)}_t[L,H^r) \tag{5.41}$$

*where* $U^r(\cdot)$ *is the renewal function associated with the patience time distribution* $G^r$.

(3) *Suppose that the limit*

$$\lim_{L\to H^r}\sup_{N\in\mathbb{N}}\mathbb{E}\left[\overline{\eta}^{(N)}_0(L,H^r)\right]=0 \tag{5.42}$$

*holds and, if* $H^r<\infty$, *then*

$$\lim_{L\to H^r}\sup_{N\in\mathbb{N}}\mathbb{E}\left[\int_{[0,L)}\frac{1-G^r(L)}{1-G^r(x)}\,\overline{\eta}^{(N)}_0(dx)\right]=0 \tag{5.43}$$

*is also satisfied. Then the following three properties hold.*

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

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

(b) *For every* $\psi\in\mathcal{C}_b([0,H^r)\times\mathbb{R}_+)$ *and* $T\in[0,\infty)$,

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

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

$$\lim_{\delta\to 0}\limsup_N\mathbb{E}\left[S^{(N)}_\psi(t+\delta)-S^{(N)}_\psi(t)\right]=0.$$

(c) *Given* $L<H^r$ *and any sequence of measurable subsets* $B_R\subset[0,L]$ *such that the Lebesgue measure of* $B_R$ *goes to zero as* $R\to\infty$, *we have for every* $T\in[0,\infty)$,

$$\lim_{R\to\infty}\limsup_N\mathbb{E}\left[\sup_{t\in[0,T]}\overline{A}^{(N)}_{\mathbb{1}_{B_R},\eta}(t)\right]=0. \tag{5.44}$$

*Proof.* This proposition can be proved in the same way as Lemma 5.6, Lemma 5.8 and Lemma 5.9 of [17]. Note that the renewal function used in the proofs of Lemma 5.6 and 5.8 of [17] was defined from the cumulative distribution function $G$ in [17], which was assumed to have no mass at $\infty$. However the proofs given there apply here when $G^r$ may possibly have a mass at $\infty$ and the renewal function associated with $G^r$ here is now simply given by $U^r(\cdot) = \int_0^{\cdot} \sum_{n=1}^{\infty} (g^r)^{*n}(s)\, ds$, where $(g^r)^{*n}$ is the $n$-th convolution of $g^r$ on $[0, \infty)$. □

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 $F^{\eta^{(N)}_t}(x) = \eta^{(N)}_t[0, x]$. Also let $(F^{\eta^{(N)}_t})^{-1}$ be its inverse, as defined in (1.1).

**Proposition 5.6.** *For each $N \in \mathbb{N}$, $t \geq 0$ and measurable function $h$ on $[0, H^r)$,*

$$\int_{[0,H^r)} \mathbb{1}_{[0,\chi^{(N)}(t-)]}(x) h(x) \eta^{(N)}_t(dx) = \int_0^{Q^{(N)}(t)+\iota^{(N)}(t)} h((F^{\eta^{(N)}_t})^{-1}(y)) dy, \tag{5.45}$$

*where*

$$\iota^{(N)}(t) \doteq \begin{cases} 0 & \text{if } (\chi^{(N)}(t-) - \chi^{(N)}(t))(K^{(N)}(t) - K^{(N)}(t-)) = 0, \\ 1 & \text{if } (\chi^{(N)}(t-) - \chi^{(N)}(t))(K^{(N)}(t) - K^{(N)}(t-)) > 0. \end{cases} \tag{5.46}$$

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

$$\begin{aligned} &\int_{[0,H^r)} \mathbb{1}_{[0,\chi^{(N)}(t-)]}(x) h(x) \eta^{(N)}_t(dx) \\ &\quad = \sum_{j=-\mathcal{E}^{(N)}_0+1}^{E^{(N)}(t)} h\left(w^{(N)}_j(t)\right) \mathbb{1}_{\{w^{(N)}_j(t) \leq \chi^{(N)}(t-)\}} \mathbb{1}_{\{w^{(N)}_j(t) < r_j\}}. \end{aligned} \tag{5.47}$$

Moreover, by (2.6),

$$Q^{(N)}(t) = \eta^{(N)}_t[0, \chi^{(N)}(t)] = \sum_{j=-\mathcal{E}^{(N)}_0+1}^{E^{(N)}(t)} \mathbb{1}_{\{w^{(N)}_j(t) \leq \chi^{(N)}(t)\}} \mathbb{1}_{\{w^{(N)}_j(t) < r_j\}}.$$

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 $i = 1, 2, \cdots, Q^{(N)}(t)$, $(F^{\eta^{(N)}_t})^{-1}(i)$ is exactly the potential waiting time at $t$ of the $i$th customer.

Suppose that $(\chi^{(N)}(t-) - \chi^{(N)}(t))(K^{(N)}(t) - K^{(N)}(t-)) = 0$. This implies that either we have $\chi^{(N)}(t-) = \chi^{(N)}(t)$ holds or we have $\chi^{(N)}(t-) > \chi^{(N)}(t)$ and $K^{(N)}(t) = K^{(N)}(t-)$ hold, which means the head-of-the-line customer right before time $t$ reneges at time $t$. In this case, the right-hand-side of (5.47) can be re-expressed as

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

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 righthand side of (5.47) but not in $Q^{(N)}(t)$. Hence the right-hand-side of (5.47) should be re-expressed as

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

□

As an immediate consequence of (2.25), Lemma 5.4, (5.32) and Proposition 5.6, we obtain the following alternative representation for the compensator $A^{(N)}_{\theta^{(N)},\eta}$ of $R^{(N)}$:

$$A^{(N)}_{\theta^{(N)},\eta}(t) \doteq \int_0^t \left( \int_0^{Q^{(N)}(t)+\iota^{(N)}(t)} h^r((F^{\eta_s^{(N)}})^{-1}(y))dy \right) ds, \qquad t \in [0,\infty), \tag{5.48}$$

where $\iota^{(N)}$ is given by (5.46).

## 6. Tightness of Pre-limit Sequences

The main objective of this section is to show that, under suitable assumptions, the sequences of scaled state processes $\{(\overline{X}^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}^{(N)})\}$ and auxiliary processes are tight. Specifically, from (2.23) and (5.18) it is clear that for every $t$, $\overline{D}^{(N)}_\cdot(t) : \varphi \mapsto \overline{D}^{(N)}_\varphi(t)$ and $\overline{A}^{(N)}_{\cdot,\nu} : \varphi \mapsto \overline{A}^{(N)}_{\varphi,\nu}(t)$ are Radon measures on $[0, H^s)$ and, likewise from (2.24) and (5.31) it follows that $\overline{S}^{(N)}_\cdot(t) : \psi \mapsto \overline{S}^{(N)}_\psi(t)$ and $\overline{A}^{(N)}_{\cdot,\eta} : \psi \mapsto \overline{A}^{(N)}_{\psi,\eta}(t)$ define Radon measures on $[0, H^r)$. It is also easy to see that these processes are càdlàg. Now, define

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

equipped with the product metric. Then $\mathcal{Y}$ is clearly a Polish space. Also, let

$$\begin{aligned}\overline{Y}^{(N)} \doteq\ & \Big(\overline{X}^{(N)}(0), \overline{E}^{(N)}, \overline{X}^{(N)}, \overline{R}^{(N)}, \overline{\nu}_0^{(N)}, \overline{\nu}^{(N)}, \overline{\eta}_0^{(N)}, \overline{\eta}^{(N)}, \\ & \overline{A}^{(N)}_{\cdot,\nu}, \overline{D}^{(N)}_\cdot, \overline{A}^{(N)}_{\cdot,\eta}, \overline{S}^{(N)}_\cdot\Big), \qquad N \in \mathbb{N}.\end{aligned} \tag{6.49}$$

The main result of this section is

**Theorem 6.1.** *Suppose Assumption 3.1 is satisfied. Then the sequence $\{\overline{Y}^{(N)}\}$ defined in (6.49) is relatively compact in the Polish space $\mathcal{Y}$, and 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' criteria (see Theorem 3.8.6 of [8] for details) for the relative compactness of a sequence $\{\overline{H}^{(N)}\}$ of processes in $\mathcal{D}_{\mathbb{R}_+}[0,\infty)$.

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

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

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

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

$$\lim_{\delta\to 0} \sup_N \mathbb{E}\left[\left|\overline{H}^{(N)}(t+\delta) - \overline{H}^{(N)}(t)\right|^{\beta}\right] = 0.$$

**Lemma 6.3.** *Suppose Assumption 3.1 holds. Then the sequences $\{\overline{X}^{(N)}\}, \{\overline{D}^{(N)}\}$, $\{\overline{K}^{(N)}\}$, $\{\overline{R}^{(N)}\}, \{\overline{S}^{(N)}\}, \{\langle \mathbf{1}, \overline{\nu}^{(N)}\rangle\}, \{\langle \mathbf{1}, \overline{\eta}^{(N)}\rangle\}$, the sequences $\{D_\varphi^{(N)}\}$, $\{\overline{A}_{\varphi,\nu}^{(N)}\}$ for $\varphi \in \mathcal{C}_b([0, H^s) \times \mathbb{R}_+)$, 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, (2.19) and (3.36) that for $\varphi \in \mathcal{C}_b([0, H^s) \times \mathbb{R}_+)$,

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

Similarly, by Lemma 5.2, (2.21) and (3.36), we have for every $\psi \in \mathcal{C}_b([0, H^r) \times \mathbb{R}_+)$,

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

Together with property 3b of Proposition 5.1 and properties 2 and 3b of Proposition 5.5, we conclude that $\{\overline{D}_\varphi^{(N)}\}$, $\{\overline{S}_\psi^{(N)}\}$, $\{\overline{A}_{\varphi,\nu}^{(N)}\}$, $\{\overline{A}_{\psi,\eta}^{(N)}\}$ and $\{\overline{R}^{(N)}\}$ satisfy K1 and K2 of Proposition 6.2, and thus are relatively compact. Since $\overline{D}^{(N)} = \overline{D}_{\mathbf{1}}^{(N)}$ and $\overline{S}^{(N)} = \overline{S}_{\mathbf{1}}^{(N)}$, then $\{\overline{D}^{(N)}\}$ 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 $t \geq 0$, $\langle \mathbf{1}, \overline{\nu}_t^{(N)}\rangle \leq \overline{X}^{(N)}(t) \leq \overline{X}^{(N)}(0) + \overline{E}^{(N)}(t)$ by (2.17) and (2.12), we infer 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

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

and by (2.17) and the Lipschitz continuity of the function $[1-x]^+$ with Lipschitz constant 1, we have

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

Hence $\{\overline{X}^{(N)}\}$ 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, (2.13) implies that for every $s, t \in [0, \infty)$,

$$\left|\langle \mathbf{1}, \overline{\eta}_t^{(N)}\rangle - \langle \mathbf{1}, \overline{\eta}_s^{(N)}\rangle\right| \leq \left|\overline{E}^{(N)}(t) - \overline{E}^{(N)}(s)\right| + \left|\overline{S}^{(N)}(t) - \overline{S}^{(N)}(s)\right|, \tag{6.50}$$

$$\langle \mathbf{1}, \overline{\eta}_t^{(N)}\rangle \leq \langle \mathbf{1}, \overline{\eta}_0^{(N)}\rangle + \overline{E}^{(N)}(t). \tag{6.51}$$

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

**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 (6.51) that for $\delta \in (0,1)$,

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

Hence, we have

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

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

$$\sup_N \mathbb{E}\left[ \langle \mathbf{1}, \overline{\eta}_t^{(N)}\rangle \right] \le \sup_N \mathbb{E}\left[ \langle \mathbf{1}, \overline{\eta}_0^{(N)}\rangle + \overline{E}^{(N)}(t) \right] < \infty. \tag{6.52}$$

Then (6.52) implies

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

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 (6.52), we also have, for every $t \in [0,\infty)$,

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

and

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

This implies that $\left\{ \int_0^\cdot \langle f', \overline{\eta}_s^{(N)}\rangle ds \right\}$ and $\left\{ \int_0^\cdot \langle f', \overline{\nu}_s^{(N)}\rangle ds \right\}$ 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 $\widetilde{C}_{T,\delta}$ of $\mathcal{M}_F[0,H)$ such that*
$$\liminf_{N\to\infty}\mathbb{P}\left(\overline{\nu}_t^{(N)}\in\widetilde{C}_{T,\delta}\text{ for all }t\in[0,T]\right)>1-\delta.$$
- **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 such that $\{\overline{\pi}^{(N)}\}$ is $\mathbb{F}$-weakly tight, i.e., 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. Moreover, (5.24), (5.25), (5.42) and (5.43) hold.*

*Proof.* By Lemma 6.4 and Remark 5.11 of [17], 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, we have that for almost every $\omega\in\Omega$, $\sup_N\overline{\nu}_0^{(N)}(\omega)[0,H^s)<\infty$. Then by Lemma A 7.5 of [14], we have that for every $\varepsilon>0$, there exists $k(\omega,\varepsilon)<\infty$ such that $\sup_N\overline{\nu}_0^{(N)}(\omega)(k(\omega,\varepsilon),H^s)<\varepsilon$. Similarly, due to (2) and (4) of Assumption 3.1, we have that 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 [14], we infer that for every $\varepsilon>0$, there exists $l(\omega,\varepsilon)<\infty$ such that $\sup_N\overline{\eta}_0^{(N)}(\omega)(l(\omega,\varepsilon),H^r)<\varepsilon$. Combining the fact, proved in Lemma 6.4 that $\langle\mathbf{1},\overline{\eta}^{(N)}\rangle$ is tight with the argument for tightness of $\overline{\nu}^{(N)}$ given in Lemma 5.12 of [17], we can also establish the tightness of $\{\overline{\eta}^{(N)}\}$. We omit the details. □

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

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

*Proof.* The proof of this lemma follows the same argument as that used to prove Lemma 5.13 of [17], with Lemma 5.10, Lemma 5.6, $M$, Lemma 5.8(1), and Corollary 5.5 in [17] replaced by Lemma 6.3, Proposition 5.1(2), $H^s$, Proposition 5.1(3a) and Proposition 5.1(1) of this paper to obtain the result for $\overline{D}_\cdot^{(N)}$ and $\overline{A}_{\cdot,\nu}^{(N)}$, and by Lemma 6.3, Proposition 5.5(2), $H^r$, Proposition 5.5(3a), Lemma 5.4 and Proposition 5.5(1) of this paper to obtain the result for $\overline{S}_\cdot^{(N)}$ and $\overline{A}_{\cdot,\eta}^{(N)}$. □

## 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 (6.49). Then, by Assumption 3.1, Theorem 6.1 and the facts that $\overline{M}^{(N)}_{\cdot,\nu} = \overline{D}^{(N)}_{\cdot} - \overline{A}^{(N)}_{\cdot,\nu} \Rightarrow 0$ by Proposition 5.1(1) and $\overline{M}^{(N)}_{\cdot,\eta} = \overline{S}^{(N)}_{\cdot} - \overline{A}^{(N)}_{\cdot,\eta} \Rightarrow 0$ by Lemma 5.4, 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

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

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., $\overline{Y}^{(N)} \to \overline{Y}$. 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. We immediately obtain that, as $N \to \infty$, $\overline{D}^{(N)} = \overline{D}^{(N)}_{\mathbf{1}} \to \overline{A}_{\mathbf{1},\nu}$. Together with (2.20) and (2.12), this implies that

$$\overline{X} = \overline{X}(0) + \overline{E} - \overline{A}_{\mathbf{1},\overline{\nu}} - \overline{R}. \tag{7.53}$$

By the first part of Assumption 3.3, Lemma 5.16 and Proposition 5.17 of [17], It follows that

$$\overline{A}_{\mathbf{1},\overline{\nu}} = \int_0^{\cdot} \langle h^s, \overline{\nu}_s\rangle ds. \tag{7.54}$$

On comparing (7.53) and (7.54) with (3.45), it is clear that in order to prove Theorem 7.1, it is necessary to show that $\overline{R}$, $\overline{Q}$, and $\overline{\eta}$ satisfy the relation (3.45). This is established in Proposition 7.2 by using the representation (5.48) of the compensator of $R^{(N)}$ to determine the limit of the sequence $\{\overline{R}^{(N)}\}$ of the scaled reneging processes. The additional arguments required to complete the proof of Theorem 7.1 are provided at the end of the section.

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

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\overline{A}^{(N)}_{\theta^{(N)},\eta}(t) - \int_0^t \left(\int_0^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y))dy\right) ds\right|\right] \to 0. \tag{7.55}$$

*Moreover, $\overline{R}^{(N)} \to \overline{R}$, where*

$$\overline{R}(t) = \int_0^t \left(\int_0^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y))dy\right) ds,\ t \in [0,\infty). \tag{7.56}$$

Let $\tilde{R}(t)$ be defined by the right-hand-side of (7.56) for $t \in [0,\infty)$. Then (7.55) implies $A^{(N)}_{\theta^{(N)},\eta} \Rightarrow \tilde{R}$. Since $\tilde{R}$ is continuous, $\overline{R}^{(N)} = \overline{M}^{(N)}_{\theta^{(N)},\eta} + \overline{A}^{(N)}_{\theta^{(N)},\eta}$ by Lemma 5.4 and $\overline{M}^{(N)}_{\theta^{(N)},\eta} \Rightarrow 0$ by Proposition 5.5(1), it follows that $\overline{R}^{(N)} \Rightarrow \tilde{R}$. This implies $\tilde{R} = \overline{R}$ and thus the second statement of Proposition 7.2 follows from the first statement. To establish (7.55), first note that, using (5.48) and the elementary equality $(F^{\eta^{(N)}_s})^{-1}(N\cdot) = (F^{\overline{\eta}^{(N)}_s})^{-1}(\cdot)$, simple algebraic manipulations show that

$$\overline{A}^{(N)}_{\theta^{(N)},\eta}(t) \doteq \int_0^t \left( \int_0^{\overline{Q}^{(N)}(t)+\bar{\iota}^{(N)}(t)} h^r((F^{\overline{\eta}^{(N)}_s})^{-1}(y))dy \right) ds, \qquad t \in [0,\infty), \tag{7.57}$$

where as usual $\bar{\iota}^{(N)} \doteq \iota^{(N)}/N$ and $\iota^{(N)}$ is given by (5.46). Next, observe that for all $t \in [0,T]$ and $L \in [0, H^r)$,

$$\left|\overline{A}^{(N)}_{\theta^{(N)},\eta}(t) - \tilde{R}(t)\right| \le \overline{C}^{(N)}_1(t,L) + \overline{C}^{(N)}_2(t,L) + \overline{C}_3(t,L), \tag{7.58}$$

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

$$\begin{aligned} \overline{C}^{(N)}_1(t,L) \quad \doteq \quad & \left| \int_0^t \left( \int_0^{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s))\wedge F^{\overline{\eta}^{(N)}_s}(L)} h^r((F^{\overline{\eta}^{(N)}_s})^{-1}(y))dy \right) ds \right. \\ & \left. - \int_0^t \left( \int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h^r((\overline{F}^{\overline{\eta}_s})^{-1}(y))dy \right) ds \right|, \end{aligned} \tag{7.59}$$

$$\overline{C}^{(N)}_2(t,L) \doteq \left| \int_0^t \left( \int_{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s))\wedge F^{\overline{\eta}^{(N)}_s}(L)}^{\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s)} h^r((F^{\overline{\eta}^{(N)}_s})^{-1}(y))dy \right) ds \right|, \tag{7.60}$$

and

$$\overline{C}_3(t,L) \doteq \int_0^t \left( \int_{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)}^{\overline{Q}(s)} h^r((F^{\overline{\eta}_s})^{-1}(y))dy \right) ds. \tag{7.61}$$

As a precursor to the proof of (7.55) of Proposition 7.2, we first establish some path properties of the limiting queue measure $\overline{\eta}$ in Lemma 7.3 and some estimates in Lemma 7.4. These two preliminary results will be used in Lemma 7.5 to show that for any $L \in [0, H^r)$, $\lim_{N\to\infty} \sup_{t\in[0,T]} \left|\overline{C}^{(N)}_1(t,L)\right| = 0$ 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 $\varepsilon$.*

*Proof.* By (4.6) of Corollary 4.2, it follows that for every $L \in [0, H^r)$,

$$\begin{aligned} \overline{\eta}_t(\{L\}) \quad = \quad & \int_{[0,H^r)} \mathbb{1}_{\{L\}}(x+t) \frac{1-G^r(x+t)}{1-G^r(x)}\, \overline{\eta}_0(dx) \\ & + \int_0^t \mathbb{1}_{\{L\}}(t-s)(1-G^r(t-s))\, d\overline{E}(s). \end{aligned} \tag{7.62}$$

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 points of both $\overline{\eta}_0$ and $\overline{E}$ are countable, $\overline{\eta}_t$ is continuous at $L$ for almost every $t \ge 0$.

Next, suppose $\overline{\eta}_t(\{L\}) > 0$. Then by (7.62), at least one of the following two cases must occur:

$$\int_{[0,H^r)} \mathbb{1}_{\{L\}}(x+t) \frac{1-G^r(x+t)}{1-G^r(x)} \, \overline{\eta}_0(dx) > 0 \tag{7.63}$$

or

$$\int_0^t \mathbb{1}_{\{L\}}(t-s)(1-G^r(t-s)) \, d\overline{E}(s) > 0. \tag{7.64}$$

If (7.63) 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 $\varepsilon > 0$,

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

However, similar to (7.62), an application of (4.6) of Corollary 4.2 with $f = \mathbb{1}_{(L,L+\varepsilon)}$ shows that the term on the left-hand-side of the last display is dominated by $\overline{\eta}_t(L, L+\varepsilon)$, and so the lemma is established in this case. On the other hand, suppose (7.64) holds. The proof in this case follows a similar argument. Indeed, first note that (7.64) implies $t-L > 0$, $1-G^r(t-L) > 0$ 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 (4.6) of Corollary 4.2 then shows that

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

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

$$\left| \int_0^T \langle \ell, \overline{\eta}_s \rangle \, ds \right| \le \tilde{L}(m,T) \int_{[0,H^r)} |\ell(x)| dx. \tag{7.65}$$

(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.,*

$$\sup_N \sup_{s\in[0,T]} \int_0^L h(x) \overline{\eta}^{(N)}_s(dx) \le \tilde{C}^h_L \sup_N \left( \langle \mathbf{1}, \overline{\eta}^{(N)}_0 \rangle + \overline{E}^{(N)}(T) \right) < \infty. \tag{7.66}$$

*Proof.* The first estimate can be proved in the same manner as Lemma 5.16 of [17] due to the similarity between the dynamics of $\overline{\nu}$ and $\overline{\eta}$. The second estimate follows from (2.13) and Assumption 3.1. □

**Lemma 7.5.** *Let $L \in [0, H^r)$ and $T \geq 0$. For every continuous function $h$ on $[0,\infty)$, as $N \to \infty$,*

$$\sup_{t\in[0,T]} \left| \int_0^t \left( \int_0^{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))dy \right) ds \right.$$
$$\left. - \int_0^t \left( \int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((\overline{F}^{\overline{\eta}_s})^{-1}(y))dy \right) ds \right| \to 0. \tag{7.67}$$

*Proof.* From the convergence of $\overline{\eta}^{(N)}$ to $\overline{\eta}$ and $\overline{Q}^{(N)}$ to $\overline{Q}$, it follows that, as $N \to \infty$, $\overline{\eta}_s^{(N)} \Rightarrow \overline{\eta}_s$ 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 $\overline{\eta}_s^{(N)} \Rightarrow \overline{\eta}_s$ and $\overline{Q}^{(N)}(s) \to \overline{Q}(s)$ as $N \to \infty$ and $\overline{\eta}_s$ is continuous at $L$. Since $\overline{\eta}_s^{(N)} \Rightarrow \overline{\eta}_s$, we have $F^{\overline{\eta}_s^{(N)}}(x) \to F^{\overline{\eta}_s}(x)$ as $N \to \infty$ for all $x \in [0, H^r)$ except on a countable subset of $[0, H^r)$. Therefore, by Theorem 13.6.3 of [28], 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 $N \to \infty$,

$$\int_0^{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))dy \to \int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((\overline{F}^{\overline{\eta}_s})^{-1}(y))dy. \tag{7.68}$$

Observing that, since $\left|\bar{\iota}^{(N)}\right| \leq 1/N$, we have

$$(\overline{Q}^{(N)}(s) + \bar{\iota}^{(N)}(s)) \wedge F^{\overline{\eta}_s^{(N)}}(L) \to \overline{Q}(s) \wedge F^{\overline{\eta}_s}(L), \quad \text{as } N \to \infty. \tag{7.69}$$

we 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 (7.69), for all sufficiently large $N$, $(\overline{Q}^{(N)}(s) + \bar{\iota}^{(N)}(s)) \wedge F^{\overline{\eta}_s^{(N)}}(L) < F^{\overline{\eta}_s}(H^r-)$. For each $n \in \mathbb{N}$, by Theorem 11.5.1 of [28] and the continuity of $h$, we obtain for each $t < F^{\overline{\eta}_s}(H^r-)$, as $N \to \infty$,

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

Combining this with (7.69), we obtain (7.68).

**Case 2.** $\overline{Q}(s) \wedge F^{\overline{\eta}_s}(L) = F^{\overline{\eta}_s}(H^r-)$. we first claim that in this case $\overline{Q}(s) = F^{\overline{\eta}_s}(L) = F^{\overline{\eta}_s}(H^r-)$. Indeed, $F^{\overline{\eta}_s}(L) \leq F^{\overline{\eta}_s}(H^r-)$ because $F^{\overline{\eta}_s}$ is non-decreasing and $L < H^r$, while $\overline{Q}(s) \leq \overline{\eta}[0, H^r) = F^{\overline{\eta}_s}(H^r-)$ by (3.44). On the other hand, the reverse inequalities $\overline{Q}(s) \geq F^{\overline{\eta}_s}(H^r-)$ and $F^{\overline{\eta}_s}(L) \geq F^{\overline{\eta}_s}(H^r-)$ hold by the case assumption, and so the claim follows. Now, define $\bar{L} \doteq (\overline{F}^{\overline{\eta}_s})^{-1}(F^{\overline{\eta}_s}(H^r-))$. Then $\overline{L} = (\overline{F}^{\overline{\eta}_s})^{-1}(F^{\overline{\eta}_s}(L)) \leq L$ and

$$F^{\overline{\eta}_s}(\bar{L}) = F^{\overline{\eta}_s}(L) = F^{\overline{\eta}_s}(H^r-). \tag{7.70}$$

Moreover, it follows from the second assertion of Lemma 7.3 that $\bar{L}$ is a point of continuity for $\overline{\eta}_s$. By (7.70), the change of variables formula then yields

$$\int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} h((\overline{F}^{\overline{\eta}_s})^{-1}(y))dy = \int_{[0,H^r)} h(x)\overline{\eta}_s(dx) = \int_{[0,\bar{L}]} h(x)\overline{\eta}_s(dx). \tag{7.71}$$

Also, by Proposition 5.6 and another application of the change of variables formula, we have

$$\int_0^{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} h((F^{\overline{\eta}_s^{(N)}})^{-1}(y))dy = \int_{[0,\chi^{(N)}(s-)]} \mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx). \tag{7.72}$$

Expanding the term on the right-hand-side, we obtain

$$\begin{aligned} &\int_{[0,\chi^{(N)}(s-)]} \mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx) \\ = &\int_{[0,\bar{L}]} \mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx) - \int_{(\chi^{(N)}(s-)\wedge\bar{L},\chi^{(N)}(s-)]} \mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx) \\ &+ \int_{(\chi^{(N)}(s-)\wedge\bar{L},\bar{L}]} \mathbb{1}_{[0,L]}(x)h(x)\overline{\eta}_s^{(N)}(dx). \end{aligned} \tag{7.73}$$

Combining the last four displays, it is clear that to prove (7.68) it suffices to show that the second and the third terms on the right-hand-side of (7.73) converge to zero, as $N \to \infty$. Recall the constant $\tilde{C}_L^h$ defined in Lemma 7.4. $\tilde{C}_L^h < \infty$ since $h$ is continuous. Then the second term is bounded above by $\tilde{C}_L^h \overline{\eta}_s^{(N)}(\chi^{(N)}(s-) \wedge \bar{L}, \chi^{(N)}(s-)]$, and by Portmanteau's theorem, the fact that $\bar{L}$ is a point of continuity of $\overline{\eta}_s$ and the claim $F^{\overline{\eta}_s}(L) = F^{\overline{\eta}_s}(H^r-)$ proved above, it follows that

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

On the other hand, the third term on the right-hand-side of (7.73) is bounded above by $\tilde{C}_L^h \overline{\eta}_s^{(N)}(\chi^{(N)}(s-) \wedge \bar{L}, \bar{L}]$ and this converges to zero as $N \to \infty$ because, as shown below, $\liminf_{N\to\infty} \chi^{(N)}(s-) \geq \bar{L}$. We argue by contradiction to justify this last assertion. Suppose this assertion were false. Then there must exist a subsequence $\{N_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \chi^{(N_k)}(s-) = \bar{L} - \delta$ for some $\delta > 0$. Hence, for $k$ large enough, $\chi^{(N_k)}(s-) < \bar{L} - \delta/2$. By Lemma A.2, we have $\chi^{(N_k)}(s-) \geq \chi^{(N_k)}(s)$. Hence $\overline{\eta}_s^{(N_k)}[0, \bar{L} - \delta/2] \geq \overline{Q}^{(N_k)}(s)$ by (2.6). Sending $k \to \infty$ and using the convergence $\overline{\eta}_s^{(N_k)} \Rightarrow \overline{\eta}_s$, the fact that $[0, \bar{L} - \delta/2]$ is closed and Portmanteau's theorem, we obtain $\overline{\eta}_s[0, \bar{L} - \delta/2] \geq \overline{Q}(s)$. This contradicts the definition of $\bar{L}$, and hence completes the proof of (7.68).

Finally, we deduce (7.67) from (7.68) using the bounded convergence theorem, whose application is justified by the bounds (7.71), (7.72) and the estimate (7.66). □

We now generalize Lemma 7.5 to allow for generally locally integrable function $h^r$.

**Lemma 7.6.** *Let $L < H^r$, for every $t \in [0,\infty)$ and $N \in \mathbb{N}$, let $\overline{C}_1^{(N)}(t, L)$ be defined as in (7.59). Then for every $T \in [0,\infty)$,*

$$\lim_{N\to\infty} \sup_{t\in[0,T]} \overline{C}_1^{(N)}(t, L) = 0. \tag{7.74}$$

*Proof.* Since $h^r$ lies in $\mathcal{L}^1_{loc}[0, H^r)$ 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 $n \in \mathbb{N}$, since

(7.74) holds with $h_n^r$ in place of $h^r$ due to Lemma 7.5, in order to prove (7.74), it suffices to show that, as $n \to \infty$, the following two convergence results hold.

(7.75)
$$\sup_N \int_0^T \left( \int_0^{(\overline{Q}^{(N)}(s)+\bar{\iota}^{(N)}(s))\wedge F^{\overline{\eta}_s^{(N)}}(L)} \left| h_n^r((F^{\overline{\eta}_s^{(N)}})^{-1}(y)) - h^r((F^{\overline{\eta}_s^{(N)}})^{-1}(y)) \right| dy \right) ds \to 0,$$

and

$$\int_0^T \left( \int_0^{\overline{Q}(s)\wedge F^{\overline{\eta}_s}(L)} \left| h_n^r((F^{\overline{\eta}_s})^{-1}(y)) - h^r((F^{\overline{\eta}_s})^{-1}(y)) \right| dy \right) ds \to 0. \tag{7.76}$$

We first consider (7.75). It is easy to see that, by Proposition 5.6, for every $N, n \in \mathbb{N}$,

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

By the same argument that is used to prove Proposition 5.7 of [17], we can show that

$$\begin{aligned}
&\int_0^T \left( \int_{[0,L]} |h_n^r(x) - h^r(x)| \overline{\eta}_s^{(N)}(dx) \right) ds \\
&\qquad \le \left( \left\langle 1, \overline{\eta}_0^{(N)} \right\rangle + \overline{E}^{(N)}(t) \right) \int_0^L |h_n^r(x) - h^r(x)| dx.
\end{aligned} \tag{7.77}$$

Since $\sup_N \left( \left\langle 1, \overline{\eta}_0^{(N)} \right\rangle + \overline{E}^{(N)}(t) \right) < \infty$ due to Assumption 3.1, and $h_n^r$ converges in $\mathcal{L}^1_{loc}[0, H^r)$ to $h^r$, we obtain (7.75) from (7.77). Similarly, observe that

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

By (7.65) and the convergence of $h_n^r$ to $h^r$ in $\mathcal{L}^1_{loc}[0, H^r)$, the term on the right-hand-side of the above display converges to 0, as $n \to \infty$, and (7.76) follows. □

**Proof of Proposition 7.2.** We first show that, for any $L < H^r$,

$$\mathbb{E}\left[ \sup_{t\in[0,T]} \overline{C}_1^{(N)}(t, L) \right] \to 0 \qquad \text{as } N \to \infty. \tag{7.78}$$

Indeed, this follows from Lemma 7.6 and an application of the dominated convergence theorem, where the latter is justified because, by (7.71), (7.72) of Lemma

7.5,

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

which is bounded uniformly in $N$ by (7.66) and Assumption 3.1.

Taking first the supremum over $t\in[0,T]$, and next expectations and then the limsup of both sides of (7.58), due to (7.78) we conclude that for every $L<H^r$,

$$\begin{aligned}&\limsup_{N\to\infty}\mathbb{E}\left[\sup_{t\in[0,T]}\left|\overline{A}^{(N)}_{\theta^{(N)},\eta}(t)-\tilde{R}(t)\right|\right]\\ &\leq \sup_N\mathbb{E}\left[\overline{C}_2^{(N)}(T,L)\right]+\mathbb{E}\left[\overline{C}_3(T,L)\right].\end{aligned}\tag{7.79}$$

However, by Proposition 5.5(3a), as $L\to H^r$,

$$\sup_N\mathbb{E}\left[\overline{C}_2^{(N)}(T,L)\right]\leq\sup_N\mathbb{E}\left[\int_0^T\int_{[L,H^r)}h^r(x)\overline{\eta}_s^{(N)}(dx)ds\right]\to 0.$$

On the other hand, by the same argument that was used to show $\mathbb{E}[R_3(m)]\to 0$ as $m\to M$ in the proof of Proposition 5.17 in [17], it follows that Assumption 3.3 implies

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

Then the proposition is a direct consequence of the last three displays and the fact that $\overline{R}^{(N)}\to\overline{R}$.

Combining the above results, we now prove the main limit result.

**Proof of Theorem 7.1.** Fix $t\in[0,\infty)$ such that $\overline{\nu}_t^{(N)}\xrightarrow{w}\overline{\nu}_t$, $\overline{\eta}^{(N)}\xrightarrow{w}\overline{\eta}$, $\overline{E}^{(N)}(t)\to\overline{E}(t)$, $\overline{X}^{(N)}(t)\to\overline{X}(t)$, $\overline{R}^{(N)}(t)\to\overline{R}(t)$, $\overline{A}^{(N)}_{\cdot,\nu}\xrightarrow{w}\overline{A}_{\cdot,\nu}$, $\overline{D}^{(N)}_{\cdot}\xrightarrow{w}\overline{A}_{\cdot,\nu}$, $\overline{A}^{(N)}_{\cdot,\eta}\xrightarrow{w}\overline{A}_{\cdot,\eta}$, $\overline{S}^{(N)}_{\cdot}\xrightarrow{w}\overline{A}_{\cdot,\eta}$ as $N\to\infty$. Since $\overline{Y}^{(N)}\to\overline{Y}$ a.s., this occurs for $t$ outside a countable set. By Proposition 5.17 of [17], this implies that as $N\to\infty$,

$$\overline{D}^{(N)}_{\varphi}(t)\to\overline{A}_{\varphi,\nu}(t)=\int_0^t\langle\varphi(\cdot,s)h^s(\cdot,s),\overline{\nu}_s\rangle\,ds,\qquad\varphi\in\mathcal{C}_b([0,H^s)\times\mathbb{R}_+).\tag{7.80}$$

An analogous argument of Proposition 5.17 of [17] also implies that as $N\to\infty$,

$$\overline{S}^{(N)}_{\psi}(t)\to\overline{A}_{\psi,\eta}(t)=\int_0^t\langle\psi(\cdot,s)h^r(\cdot,s),\overline{\eta}_s\rangle\,ds,\qquad\psi\in\mathcal{C}_b([0,H^r)\times\mathbb{R}_+).\tag{7.81}$$

In particular, when $\varphi=\psi=\mathbf{1}$, the above two displays together with Proposition 5.1(2) and Proposition 5.5(2) imply that (3.39) 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), (2.17), (2.14), (2.20), (2.12), (2.6), (7.56), this implies that all the equations in Definition 3.3 are satisfied at time $t$ expect (3.40) and (3.42).

It only remains to show that (3.40) and (3.42) are also satisfied at time $t$. This can be proved first for (3.40) and then for (3.42) by using the argument in proving that $\overline{\nu}$ and $\overline{K}$ satisfy (3.5) in the proof of Theorem 5.15 in [17]. 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. Observe that the virtual waiting time defined in (2.18) can be rewritten in terms of the fluid-scaled quantities as

$$W^{(N)}(t) \doteq \inf\left\{s \geq 0 :\ \overline{D}^{(N)}(t+s) - \overline{D}^{(N)}(t) + \overline{\mathcal{T}}_t^{(N)}(s) > \overline{Q}^{(N)}(t)\right\}. \tag{7.82}$$

We first show that for each $t \in [0,\infty)$, $\overline{\mathcal{T}}_t^{(N)} \Rightarrow \overline{\mathcal{T}}_t$ as $N \to \infty$. 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 (that is, before entering service) in the period $[t, t+s]$, there must exist a time $u \in [t, t+s]$ when the customer is still in queue (i.e., has not yet entered service) or, equivalently, satisfies $w_j^{(N)}(u) < \chi^{(N)}(u-)$, the waiting time of the head-of-the-line customer just prior to $u$, and the customer reneges, so that her potential waiting time changes from linear increase to being flat. Therefore, for each $s \in [0,\infty)$, $\mathcal{T}_t^{(N)}(s)$ can be alternately expressed as

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

Let

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

and

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

It is easy to see that $\mathcal{T}_t^{(N)}(s) = \mathcal{T}_t^{(N),1}(s) - \mathcal{T}_t^{(N),2}(s)$, $\mathcal{T}_t^{(N),1}(s) = R^{(N)}(t+s) - R^{(N)}(t)$ and $\mathcal{T}_t^{(N),2}(s) \leq S^{(N)}(t+s) - S^{(N)}(t)$. Therefore, an application of Kurtz' 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 $\overline{\mathcal{T}}_t^{(N)}(s) \Rightarrow \overline{\mathcal{T}}_t$ 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 (7.54), $\overline{A}_{\mathbf{1},\overline{\nu}} = \int_0^{\cdot}\langle h^s, \overline{\nu}_s\rangle ds$ 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 $T \in [0,\infty)$, $\sup_{s\in[0,T]} |\overline{D}^{(N)}(t+s) - \overline{A}_{\mathbf{1},\overline{\nu}}(t+s)| \to 0$ and $\sup_{s\in[0,T]} |\overline{\mathcal{T}}_t^{(N)}(s) - \overline{\mathcal{T}}_t| \to 0$ as $N \to \infty$. From (7.82), 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 [23] and the assumption that $\overline{A}_{\mathbf{1},\overline{\nu}}$ is uniformly strictly increasing, we also know that $(\overline{D}^{(N)})^{-1} \to \overline{A}^{-1}_{\mathbf{1},\overline{\nu}}$ 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 (7.82) and the right-continuity of $\overline{D}^{(N)}, \overline{Q}^{(N)}$ and $\overline{\mathcal{T}}^{(N)}_t$, we then have $\overline{D}^{(N_n)}(t+\overline{W}^{(N_n)}(t)) - \overline{D}^{(N_n)}(t) + \overline{\mathcal{T}}^{(N_n)}_t(\overline{W}^{(N_n)}(t)) \ge \overline{Q}^{(N_n)}(t)$. Sending $n \to \infty$, we obtain

$$\overline{A}_{\mathbf{1},\overline{\nu}}(t+W^*) - \overline{A}_{\mathbf{1},\overline{\nu}}(t) + \overline{\mathcal{T}}_t(W^*) \ge \overline{Q}(t). \tag{7.83}$$

Together with (3.57), 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 $\overline{A}_{\mathbf{1},\overline{\nu}}$ is uniformly strictly increasing and $\overline{\mathcal{T}}_t$ is non-decreasing, the inequality (7.83) implies that $\overline{A}_{\mathbf{1},\overline{\nu}}(t+w) - \overline{A}_{\mathbf{1},\overline{\nu}}(t) + \overline{\mathcal{T}}_t(w) > \overline{Q}(t)$. Therefore, for sufficiently large $N$, we have $\overline{D}^{(N)}(t+w) - \overline{D}^{(N)}(t) + \overline{\mathcal{T}}^{(N)}_t(w) > \overline{Q}^{(N)}(t)$ 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.

## References


[1] S. Asmussen, *Applied probability and queues*, 2nd edition ed., Springer-Verlag, New York, 2003.

[2] F. Baccelli and G. Hebuterne, On queues with impatient customers. In *Performance '81*, ed. E. Gelenbe (North-Holland Publ. Cy., Amsterdam), pp 159-179, 1981.

[3] A. Bassamboo, J.M. Harrison and A. Zeevi. Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, **51** 249–285, 2005.

[4] O.J. Boxma and P.R. de Waal, Multiserver queues with impatient customers, *Proceedings of ITC*, **14**, 1994.

[5] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, *Statistical analysis of a telephone call center: a queueing science perspective*, JASA **100** (2005), no. 469, 36–50.

[6] D. Dawson, B. Maisonneuve, and J. Spencer, *Ecole d'eté de probabilités de Saint-Flour XXI*, Lecture Notes in Mathematics, vol. 1541, Springer-Verlag, Berlin, 1991.

[7] P. Dupuis and R. Ellis, *A weak convergence approach to the theory of large deviations*, John Wiley & Sons, New York, 1997.

[8] S.N. Ethier and T.G. Kurtz, *Markov processes: Characterization and convergence*, Wiley, 1986.

[9] N. Gans, G. Koole and A. Mandelbaum. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.

[10] O. Garnett, A. Mandelbaum and M.I. Reiman. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 3:208–227, 2002.

[11] J.M. Harrison A method for staffing large call centers based on stochastic fluid models, *Manufacturing and Service Operations Management*, **7** 1:20–36, 2005.

[12] J. Jacod and A.N. Shiryaev, *Limit theorems for stochastic processes*, Springer-Verlag, Berlin, 1987.

[13] A. Jakubowski, *On the Skorokhod topology*, Ann. Inst. H. Poincaré **B22** 263–285, 1986.

[14] O. Kallenberg, *Random Measures*, Academic Press, 1975.

[15] W. Kang and K. Ramanan. Long-time behavior of fluid limits of many-server queues with reneging. *Working paper*, 2008.

[16] W. Kang and K. Ramanan. Functional central limits for many-server queues with reneging. *Working paper*, 2008.
[17] H. Kaspi and K. Ramanan, *Law of large numbers limits for many-server queues*, *Preprint*, 2007.
[18] H. Kaspi and K. Ramanan, *Central limit theorems for many-server queues in the QED regime*, *Working Paper*, 2008.
[19] A. Mandelbaum, W. Massey and M. Reiman, Strong approximations for Markovian service networks, *Queueing Systems*, **30**, 149–201, 1998.
[20] A. Mandelbaum and S. Zeltyn, Call centers with impatient customers: many-server asymptotics of the $M/M/N+G$ queue. *QUESTA* **51**, 361–402, 2005.
[21] A. Mandelbaum and S. Zeltyn, Staffing many-server queues with impatient customers: constraint satisfaction in call centers *Preprint*, 2008.
[22] K.R. Parthasarathy, *Probability measures on metric spaces*, Academic Press, 1967.
[23] K. Ramanan and M. Reiman, *Fluid and heavy traffic diffusion limits for a generalized processor sharing model*, Ann. Appl. Prob. **13** (2003), no. 1, 100–139.
[24] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales, Vol 1: Foundations* (1994), Cambridge University Press, Cambridge, U. K.
[25] L. C. G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales, Vol 2: Itô Calculus* (1994), Cambridge University Press, Cambridge, U. K.
[26] W. Whitt, *Fluid models for multiserver queues with abandonments*, Oper. Res. **54** (2006), no. 1, 37–54.
[27] ———, *Martingale proofs of many-server heavy traffic limits*, Working Paper, 2007.
[28] ———, *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*, Springer, 2002.

## Appendix A. Explicit Construction of the State Processes

In this section, we construct all state processes and auxiliary processes in Section 2.2 from the initial data $\{\mathcal{E}_0^{(N)}, X^{(N)}(0), w_j^{(N)}(0), a_j^{(N)}(0),\ j = -\mathcal{E}_0^{(N)}+1, \cdots, 0\}$, $\{\alpha_E^{(N)}(t), t \in [0,\infty)\}$, $\{v_j,\ j \in \mathbb{Z}\}$ 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 $\alpha_E$ using the relation (2.1). For $j \in \mathbb{N}$, let $w_j(0) = 0$. 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 $j = -\mathcal{E}_0 + 1, \cdots, 0, \cdots, E(\tau_\ell)$ and $t \in [\tau_\ell, \infty)$, let

$$\begin{aligned}
w_j^\ell(t) &\doteq (w_j(\tau_\ell) + t - \tau_\ell) \wedge r_j, \\
a_j^\ell(t) &\doteq \begin{cases} 0 & \text{if } w_j(\tau_\ell) = r_j \text{ or } w_j(\tau_\ell) \le \chi^\ell(\tau_\ell), \\ (a_j(\tau_\ell) + t - \tau_\ell) \wedge v_j & \text{if } \chi^\ell(\tau_\ell) < w_j(\tau_\ell) < r_j, \end{cases} \\
\eta_t^\ell &\doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \delta_{w_j(t)} \mathbb{1}_{\left\{\frac{dw_j}{dt}(t+)>0\right\}}, \\
\nu_t^\ell &\doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \delta_{a_j(t)} \mathbb{1}_{\left\{\frac{da_j}{dt}(t+)>0\right\}}, \\
R^\ell(t) &\doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{w_j(s)\le\chi^l(s-),\ \frac{dw_j}{dt}(s-)>0,\ \frac{dw_j}{dt}(s+)=0\right\}}, \\
D^\ell(t) &\doteq \sum_{j=-\mathcal{E}_0+1}^{E(\tau_\ell)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{da_j}{dt}(s-)>0,\ \frac{da_j}{dt}(s+)=0\right\}}.
\end{aligned}$$

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 $Y = w_j, a_j, j \in -\mathcal{E}_0+1, \ldots, E(\tau_\ell)$, $R$, $D$, $\eta$, $\nu$ and $\chi$ and set $Y(t) = Y(\tau_\ell)$ for $Y = X,\ Q,\ w_j, a_j, j > E(\tau_\ell)$. Moreover, define

$$\begin{aligned}
X(\tau_{\ell+1}) &\doteq X(\tau_\ell) + E(\tau_{\ell+1}) - E(\tau_\ell) - D(\tau_{\ell+1}) + D(\tau_\ell) \\
&\qquad -R(\tau_{\ell+1}) + R(\tau_\ell), \\
\eta_{\tau_{\ell+1}} &\doteq \eta^\ell_{\tau_{\ell+1}} + (E(\tau_{\ell+1}) - E(\tau_\ell))\delta_0,
\end{aligned}$$

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 the equations (A.1) and (A.2), but with $\ell$ replaced by $\ell + 1$, and the procedure can be reiterated. Now, $\max\{\ell : \tau_\ell \le t\}$ 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 the equations (2.14) and (2.13).

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

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

where $\zeta_j = \inf\{t > 0 : \ E(t) = j\}$. Hence $w_j$ constructed is indeed the potential waiting time process of customer $j$. It is also not to hard to see that $a_j$ constructed is the age process of customer $j$ and satisfies (2.7). We next show that the process $\chi$ constructed 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 can see that

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

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

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

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

This shows that the $\eta$ constructed satisfies (2.3). The similar argument shows that the processes $\nu$, $D$ and $R$ constructed satisfy (2.8), (2.9) and (2.11), respectively. Finally, $K$ and $S$ satisfy (2.14) and (2.13) by the construction.

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

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

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 $E$, $\eta$, $\nu$, $D$ and $S$ are $\mathcal{F}_t$-adapted. We next show that $\chi$ 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 $\eta_0$ 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 $\tau_\ell$ is an $\mathcal{F}_t$-stopping time. Since $\chi_t = \chi^\ell_t$ if $t \in [\tau_\ell, \tau_{\ell+1})$, then $\chi$ is $\mathcal{F}_t$-adapted. Due to equations (2.11) and (2.12), since $\{\chi^{(N)}(s-),\ s \leq t\}$ is $\mathcal{F}^{(N)}_t$-adapted, then $X^{(N)}$ is $\mathcal{F}^{(N)}_t$-adapted. Moreover, by (2.11), we have $R$ is $\mathcal{F}^{(N)}_t$-adapted. Finally, 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.** *$\chi$ 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, $\chi_t = \chi^\ell_t$ if $t \in [\tau_\ell, \tau_{\ell+1})$. Since $\chi^\ell$ is linear on $[\tau_\ell, \tau_{\ell+1})$, $\chi$ is piecewise linear. Also $\chi$ 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 $\chi$ 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 $\chi$ is càdlàg and piecewise linear. □

*E-mail address*: weikang@andrew.cmu.edu

Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA
*E-mail address*: kramanan@math.cmu.edu