# LARGE-TIME LIMIT OF NONLINEARLY COUPLED MEASURE-VALUED EQUATIONS THAT MODEL MANY-SERVER QUEUES WITH RENEGING

RAMI ATAR, WEINING KANG, HAYA KASPI, AND KAVITA RAMANAN

Abstract. The large-time behavior of a nonlinearly coupled pair of measure-valued transport equations with discontinuous boundary conditions, parameterized by a positive real-valued parameter $\lambda$, is considered. These equations describe the hydrodynamic or fluid limit of many-server queues with reneging (with traffic intensity $\lambda$), which model phenomena in diverse disciplines, including biology and operations research. For a broad class of reneging distributions with finite mean and service distributions with finite mean and hazard rate function that is either decreasing or bounded away from zero and infinity, it is shown that if the fluid equations have a unique invariant state, then the Dirac measure at this state is the unique random fixed point of the fluid equations, which implies that the stationary distributions of scaled $N$-server systems converge to the unique invariant state of the corresponding fluid equations. Moreover, when $\lambda \neq 1$, it is shown that the solution to the fluid equation starting from any initial condition converges to this unique invariant state in the large time limit. The proof techniques are different under the two sets of assumptions on the service distribution. When the hazard rate function is decreasing, a reformulation of the dynamics in terms of a certain renewal equation is used, in conjunction with recursive asymptotic estimates. When the hazard rate function is bounded away from zero and infinity, the proof uses an extended relative entropy functional as a Lyapunov function. Analogous large-time convergence results are also established for a system of coupled measure-valued equations modeling a multiclass queue.

## 1. Introduction

1.1. **Background, Motivation and Results.** The focus of this work is the analysis of the large-time behavior of a nonlinearly coupled pair of measure-valued transport equations with discontinuous boundary conditions that describe the hydrodynamic or fluid limit of a many-server queue with reneging. Many-server queues with reneging arise in a range of applications, including as models of computer networks, telephone call centers or (more general) customer contact centers [17, 30, 37], and enzymatic processing networks in biology, where reneging seeks to model the phenomenon of dilution (see, e.g., [31]). A basic model, also referred to as the GI/G/N+G queue, consists of a system with $N$ identical servers, to which jobs arrive with independent and identically distributed (i.i.d.) service requirements that are drawn from a general distribution, with each job also being equipped with an i.i.d. patience time drawn from another general distribution. Depending on the application, the servers represent processors, call center agents or enzymes, and the jobs represent packets, customers with tasks or proteins. Arriving jobs enter service immediately if there is an idle server available, else they join the back of the queue. As servers become available, jobs from the queue start service in the order of arrival. Once a job completes service, it departs the system. In addition, jobs also renege from the queue if the amount of time they have been in queue exceeds their patience time. Important system performance measures of interest include the stationary waiting time and queue distributions. In the special case when arrivals are Poisson and the service distribution is exponential, but the abandonment distribution is general, explicit formulas for the scaled steady-state distributions were

*Date*: July 12, 2021.

2010 *Mathematics Subject Classification.* Primary: 60F17, 60K25, 90B22; Secondary: 60H99, 35D99.

*Key words and phrases.* many-server queues, GI/G/N+G queue, fluid limits, reneging, abandonment, measure-valued processes, renewal equation, large-time behavior, stationary distribution, call centers, enzymatic processing networks, transport equation, age-structured population models.

KR is supported by the U.S. Army Research Office via Grant W911NF2010133 and RA is supported by the Israel Science Foundation grant 1035/20.

obtained in [10], and their asymptotics as $N$, the number of servers, goes to infinity, were studied in [42]. However, the case of general service distributions, which is relevant for many applications, is more challenging and it appears not feasible to obtain exact analytical expressions for these quantities for general service and abandonment distributions. Instead, one often resorts to obtaining asymptotic approximations that are exact in the limit as the number of servers goes to infinity.

In [23] the state of an $N$-server queue at time $t$ is represented in terms of two coupled measures, the queue measure and the server measure. The queue measure encodes jobs currently in the queue and has a unit Dirac delta mass at the amount of time elapsed since that job entered the system, whereas the server measure $\nu^N$ keeps track of jobs currently in service and has a unit Dirac delta mass at the age of each such job, where the age is the amount of time elapsed since the job entered service. For analytical purposes, it turns out that the queue measure itself is more conveniently represented in terms of a potential queue measure $\eta^N$, which keeps track of the times elapsed since entry into the system of all jobs (whether or not they have abandoned or entered service), and not only of jobs currently in the queue, as well as the total number $X^N$ of jobs in system. Under fairly general conditions on the service and patience distributions (see Assumption 2.1), it was shown in [23] that when the average arrival rate or traffic intensity converges to $\lambda > 0$, the rescaled state descriptor $(X^{(N)}, \nu^{(N)}, \eta^{(N)})$ converges to a deterministic limit $(X, \nu, \eta)$, where $\nu$ and $\eta$ are characterized as the unique weak solutions to a nonlinearly coupled system of deterministic measure-valued transport equations, subject to discontinuous boundary conditions, which we refer to as the fluid equations (see Definition 2.3).

In this work we study the long-time behavior of the solution $(X, \nu, \eta)$ to the measure-valued fluid equations obtained in [23] under Assumption 2.1 and the assumption that the fluid equations admit a unique invariant state (equivalently, fixed point). We study the subcritical, critical and supercritical regimes, characterized, respectively, by the regions where $\lambda < 1$, $\lambda = 1$ and $\lambda > 1$, additionally assuming in the subcritial and supercritical regimes that the hazard rate function of the service time distribution is either decreasing or bounded away from zero and infinity. (Here, and in the sequel, we will say decreasing to mean non-increasing.) Our main results are summarized in Theorem 3.2. Specifically, we show that when $\lambda \neq 1$, from any initial condition, the solution to the fluid equations converges to the unique invariant state in the large time limit and when $\lambda = 1$ and the hazard rate function is decreasing then the total mass of $\nu$ (which represents the mass of busy servers in the fluid system) converges to 1. In all cases above, we show that the fluid equations have a unique invariant distribution (or equivalently, unique *random* fixed point, to use a term introduced later in this paper; see Definition 2.10). This crucially implies that the stationary distributions of the $N$-server dynamics converge to the invariant state of the fluid dynamics, the proof of which was one of the motivations of this work. In particular, as elaborated in Remark 3.3, it is the uniqueness of the invariant distribution (or random fixed point) for the fluid dynamics, rather than just the uniqueness of an invariant state, that is relevant for the convergence of stationary distributions of the $N$-server dynamics. In the absence of reneging, such long-time convergence results were established for a single-class system in Proposition 6.1 of [25] for the subcritical regime and in Theorem 3.9 of [25] for the critical regime, with the latter requiring an additional finite second moment assumption. In the presence of reneging, although the system is in a sense more stable (e.g., the system is also stable in the supercritical regime, making it of particular interest), certain monotonicity properties are lost and the fluid equation dynamics are considerably more complicated, making the analysis significantly more challenging.

The proof of convergence in the subcritical regime is obtained via a direct analysis of the fluid equations (see Section 4.1). The proofs in the critical and supercritical cases are considerably more subtle, and rely on rather different arguments under the two sets of assumptions. When the hazard rate function of the service distribution is decreasing, we use a reformulation in terms of renewal equations, in conjunction with certain recursive estimates, and the convergence of the measure-valued state processes is with respect to the weak topology (see Section 4.3). These arguments are inspired by those used in the work [26], which studies the long-time behavior of fluid equations for the GI/G/N+G model under the assumption that the service time distribution has a concave or convex renewal function (which is implied by decreasing hazard rate functions). However, the fluid equations of [26] are based on a different measure-valued state representation, involving

residual service times of customers rather than ages, and moreover, convergence is established in [26] for the queue process, not the measure-valued process. Thus the results of [26] do not directly apply. Moreover, we also need to establish additional estimates to prove convergence of the measure-valued process $\nu$ in the supercritical case.

The arguments used when the hazard rate function $h^s$ of the service distribution is bounded away from zero and infinity are of a completely different nature. These results address a class of distributions not covered by related results in the literature for many-server systems. They are based on the analysis of weak solutions to partial differential equations, and entail showing that an extended relative entropy functional (that takes in arguments that are not necessarily probability measures) serves as a Lyapunov functional for the dynamics. As a result, the convergence of the measure-valued state processes is with respect to the stronger total variation topology. The analysis here is somewhat reminiscent of the study of age-structured population models arising in biology (see, e.g., [13, 36, 32]). Indeed, although the server measure $\nu$ need not have a density, and in fact will typically not have a continuous density, a purely formal derivation (see Section 4.2.1) shows that the density of $\nu$ satisfies a partial differential equation (PDE) that is similar to age-structured population models. However, our fluid equation differs crucially from such models in several aspects that make it in some ways harder to analyze. One issue, as elaborated in Remark 4.6 is that the service time hazard rate function, which appears as a coefficient in the PDE, is not integrable on $[0, \infty)$, and thus the relative entropy method developed in [32] does not apply. But, more significantly, a key complicating factor is the fact that the boundary condition for $\nu$ is discontinuous and assumes a different form depending on whether the total mass of $\nu$ is less than or equal to, 1. In fact, a significant challenge in the analysis of the critical and supercritical regimes is the to control the oscillations of the total mass of $\nu$ below 1.

Finally, we also analyze the long-time behavior of fluid equations for a multiclass model under a nonpreemptive priority policy, which was formulated in [9] and used therein to establish asymptotic optimality of the policy when the reneging distribution is exponential (see Definition 5.1). In the case that service time distribution is class-independent and satisfies the same conditions as above, reneging times are exponential, but possibly class-dependent, and the fluid equations have a unique invariant state, we establish (in Theorem 5.2) uniqueness of the random fixed points and analogous long-time convergence results in the supercritical regime. It should be mentioned here that in the queueing context, other works that have studied long-time behavior of measure-valued fluid equations using Lyapunov functionals include [35, 19, 38]. All of these works focus on the dynamics of residual times for jobs in bandwidth sharing and processor sharing models, which have a different structure from the measure-valued equations arising from our fluid equations.

An interesting open problem for future investigation would be to determine precisely the full class of service distributions for which such long-time convergence holds, and also whether there is a unified proof for all cases, at least in the supercritical regime. In addition, in the critical regime, a more complete study of the convergence of the state process even under the conditions imposed here would also be of interest. Moreover, the techniques developed here may be potentially used to establish such convergence results for more general many-server systems, including load-balancing systems with general service distributions, where the fluid limits are described in terms of a system of coupled measure-valued equations [2] or partial differential equations [5], and uniqueness of the fixed point holds under general conditions [1]. For the multiclass model, it is of interest to investigate broader conditions, such as class-dependent service distributions and less restrictive assumptions on the hazard rate, under which convergence holds. This would also allow to treat asymptotic optimality of the aforementioned index rule in broader settings.

1.2. **Ramifications for stationary distributions of $N$-server queues.** The results of this paper also shed insight into the (law-of-large-numbers) scaled limit of stationary distributions of $N$-server queues for a much broader class of service distributions. More precisely, it follows from Theorem 3.2 and Theorem 7.1 of [24] that the measure-valued state dynamics $(X^{(N)}, \eta^{(N)}, \nu^{(N)})$ for each $N$-server system describe an ergodic Feller process with a unique stationary distribution, whereas Theorem 3.3 of [24] shows that the sequence of scaled stationary distributions $\pi^{(N)}$ of the normalized state is tight. Moreover, the latter theorem also states that any subsequential limit of $\pi^{(N)}$ must coincide with the (deterministic) invariant state of the fluid

equations, whenever the latter is unique. However, there is a gap in the proof of this statement in [24]. *A priori* one only knows that any subsequential limit of the scaled stationary distributions of $N$-server queues with reneging is a random fixed point of the fluid equations (see Definition 2.10), and not that it is necessarily equal to a deterministic fixed point. However, as shown in Proposition 4.16 of the present paper, when there is convergence of the fluid equations to a unique invariant state from any initial condition (or, when $\lambda = 1$, just convergence of $\eta_t$ and the fraction of busy server servers), it follows that the set of random fixed points is in fact equal to the Dirac delta measure at the unique invariant state, thus closing the gap in Theorem 3.3 of [24].

Our work in the multi-class setting also closes an exactly analogous gap present in Theorem 4.4 of [9]. Indeed, one of the auxiliary goals of this work is to (partially) fix the gaps in these proofs, under the additional assumptions on the service distribution imposed herein (see Remark 3.3 for further elaboration of this point). In the case of [9], the gap also affects the validity of Theorem 5.1 there, regarding the asymptotic optimality of an index policy, referred to as the $c\mu/\theta$ rule, which was introduced in [7]. The results obtained in this paper validate the asymptotic optimality result, Theorem 5.1 of [9] under the additional assumption that the service time distributions do not depend on the class. (Note, however, that there is no problem with the validity of the asymptotic optimality results of the $c\mu/\theta$ rule stated in [7] and [8], which deal with the case of exponential service time distributions. Also, note recent developments on this policy under various additional settings in [27]). Finally, we note that limits of stationary distributions of many-server systems in the (so-called Halfin-Whitt) diffusive regime have been considered in [20, 3, 4] in the absence of abandonment, and in [22, 16, 21] in the presence of abandonment.

1.3. **Organization of the rest of the Paper.** In Section 2.1 we introduce the fluid equations in the single-class setting, and in Section 2.2 define their invariant states. In Section 3 we state our assumptions and the main results, and provide the proofs in Section 4. Finally, in Section 5 we introduce the multiclass fluid equations and establish our convergence results in that setting. First, in Section 1.4, we introduce common notation that is used throughout the paper.

1.4. **Common Notation and Terminology.** The following notation will be used throughout the paper. $\mathbb{Z}$ is the set of integers, $\mathbb{N}$ is the set of strictly positive integers, $\mathbb{R}$ is set of real numbers, $\mathbb{R}_+$ the set of 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 $a \vee 0$. Also, given a set $A$, we will use $1_A$ to denote the indicator function, which is 1 on $A$ and zero otherwise.

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 $\mathcal{L}^1_{loc}[0, H)$, respectively, represent the spaces of integrable and locally integrable functions on $[0, H)$, where a locally integrable function $f$ on $[0, H)$ is a measurable function on $[0, H)$ that satisfies $\int_{[0,a]} f(x)dx < \infty$ for all $a < H$. Given any càdlàg, real-valued function $f$ defined on $[0, \infty)$, we define $\|f\|_T := \sup_{s\in[0,T]} |f(s)|$ for every $T < \infty$, and let $\|f\|_\infty := \sup_{s\in[0,\infty)} |f(s)|$, which could possibly take the value $\infty$. In addition, the support of a function $f$ is denoted by $\text{supp}(f)$. Given a nondecreasing function $f$ on $[0, \infty)$, $f^{-1}$ denotes the inverse function of $f$, defined precisely as

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

For each differentiable function $f$ defined on $\mathbb{R}$, $f'$ denotes the first derivative of $f$. For each function $f(t, x)$ defined on $\mathbb{R} \times \mathbb{R}^n$, we will use both $f_x$ and $\partial_x f$ to denote the partial derivatives of $f$ with respect to $x$, and likewise, both $f_t$ and $\partial_t f$ to denote the partial derivatives of $f$ with respect to $t$. We use $\mathbf{1}$ to denote the function that is identically equal to 1. We will mostly be interested in the case when $E = [0, H)$ and $E = [0, H) \times \mathbb{R}_+$, for some $H \in (0, \infty]$. To distinguish these cases, we will usually use $\psi$ 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 $\psi$ 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.

We use $\mathcal{P}(E)$ and $\mathcal{M}(E)$ to denote, respectively, the space of Radon measures on a metric space $E$, endowed with the Borel $\sigma$-algebra, and let $\mathcal{M}_F(E)$ denote the subspace of finite measures in $\mathcal{M}(E)$, and $\mathcal{M}_F^c(E)$ the subspace of continuous measures (i.e., measures that do not charge points) in $\mathcal{M}_F(E)$. The symbol $\delta_x$ will be used to denote the measure with unit mass at the point $x$ and, with 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 $\mathcal{M}_F[0,H)$ or $\mathcal{M}_F^c[0,H)$ instead of $\mathcal{M}_F([0,H))$ or $\mathcal{M}_F^c([0,H))$, respectively. For any Borel measurable function $\psi:[0,H)\to\mathbb{R}$ that is integrable with respect to $\xi\in\mathcal{M}[0,H)$, we often use the short-hand notation

$$\langle\psi,\xi\rangle := \int_{[0,H)}\psi(x)\,\xi(dx),$$

and likewise, for any Borel measurable function $\varphi:[0,H)\times[0,\infty)\to\mathbb{R}$ and $t>0$ such that $x\mapsto\varphi(\cdot,t)$ is integrable with respect to $\xi\in\mathcal{M}[0,H)$, we often use the short-hand notation

$$\langle\varphi(\cdot,t),\xi\rangle := \int_{[0,H)}\varphi(\cdot,t)d\xi = \int_{[0,H)}\varphi(x,t)\,\xi(dx).$$

We also let $\mathcal{P}(E)$ denote the space of probability measures on $E$, equipped with the Borel $\sigma$-algebra.

For any measure $\mu\in\mathcal{M}_F[0,H)$, we define

$$F^\mu(x) := \mu[0,x], \quad x\in[0,H), \tag{1.2}$$

and we define $(F^\mu)^{-1}$ to be its right-continuous inverse:

$$(F^\mu)^{-1}(y) = \inf\{x>0 : F^\mu(x)\geq y\}. \tag{1.3}$$

Also, given $\mu,\mu_t,t\in[0,\infty)$, in $\mathcal{M}_F[0,H)$, we will use the notation $\mu_t\Rightarrow\mu$ to denote weak convergence:

$$\lim_{t\to\infty}\langle\psi,\mu_t\rangle = \langle\psi,\mu\rangle, \quad \forall\psi\in\mathcal{C}_b[0,H).$$

We will also on occasion use the total variation distance on $\mathcal{M}_F[0,H)$, denote by $d_{\mathrm{TV}}(\mu,\nu) := 2\sup_{A\in\mathcal{F}}|\mu(A)-\nu(A)|$, where $\mathcal{F}$ is the Borel $\sigma$-algebra on $[0,H)$.

Given a Polish space $\mathcal{H}$, Let $\mathcal{D}_{\mathcal{H}}[0,\infty)$ denote the space of $\mathcal{H}$-valued, càdlàg functions on $[0,\infty)$ and $\mathcal{I}_{\mathbb{R}_+}[0,\infty)$ denote the subset of non-decreasing functions $f\in\mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $f(0)=0$. Let $\mathcal{D}^+_{\mathbb{R}^J}(\mathbb{R}_+)$ denote the subset of functions in $\mathcal{D}_{\mathbb{R}^J}(\mathbb{R}_+)$ that are nonnegative and nondecreasing componentwise.

## 2. Fluid Equations and Random Fixed Points

2.1. **Fluid Equations.** We now describe the fluid equations introduced in [24]. Let $G^s$ and $G^r$ denote the cumulative distribution functions of the service time and patience time distributions, respectively. Throughout, we make the following standing assumptions on $G^s$ and $G^r$ and let $\bar{G}^s = 1-G^s$ and $\bar{G}^r = 1-G^r$ denote the corresponding complementary cumulative distribution functions. Recall that we abbreviate lower semicontinuous as lsc.

**Assumption 2.1.** *The cumulative distribution functions $G^r$ and $G^s$ satisfy $G^r(0+)=G^s(0+)=0$, and are both absolutely continuous on $[0,\infty)$ with densities $g^r$ and $g^s$ that satisfy the following properties:*

(1) *The mean patience and service times are finite: in particular,*

$$\theta^r := \int_{[0,H^r)} xg^r(x)\,dx = \int_{[0,\infty)}\bar{G}^r(x)\,dx<\infty, \tag{2.1}$$

*and, we normalize units so that*

$$\int_{[0,\infty)} xg^s(x)\,dx = \int_{[0,H^s)}\bar{G}^s(x)\,dx = 1, \tag{2.2}$$

*where*

$$H^s := \sup\{x \in [0,\infty) : G^s(x) < 1\}, \tag{2.3}$$

$$H^r := \sup\{x \in [0,\infty) : G^r(x) < 1\}, \tag{2.4}$$

*denote the right-end of the supports of the measures corresponding to $G^s$ and $G^r$, respectively.*

(2) *There exists $\bar{H}^s < H^s$ such that $h^s := g^s/\bar{G}^s$ is either bounded or lsc on $(\bar{H}^s, H^s)$, and likewise, there exists $\bar{H}^r < H^r$ such that $h^r := g^r/\bar{G}^r$ is either bounded or lsc on $(\bar{H}^r, H^r)$.*

**Remark 2.2.** Strictly speaking, $g^s$ and $g^r$ (and thus $h^s$ and $h^r$) are determined only almost everywhere. The convention implicitly adopted in the above statement is that $h^s$ (respectively, $h^r$) is almost everywhere (a.e.) equal to a function from $\mathbb{R}_+$ to itself that is bounded or lsc.

At any time $t \geq 0$, the state of the fluid system is represented by a triplet $(X(t), \nu_t, \eta_t)$, where $X(t)$ represents the total mass of jobs in system at time $t$, including those in queue and those in service, $\nu_t$ is the fluid age measure, which is a sub-probability measure on $[0, H^s)$ that assigns to any interval $(a, b)$ the (limiting) fraction of servers for whom the job currently in service has been in service for a number of time units lying in that interval, and $\eta_t$ is the fluid potential queue measure, which is a finite measure on $[0, H^r)$ that to any interval $(a, b) \subset [0, \infty)$ assigns the mass (or normalized limit) of jobs that have arrived by time $t$ and whose patience lies in that interval (irrespective of whether or not they have entered service or departed the system by time $t$). Note that the total fraction of busy servers at time $t$ is $1 - \langle \mathbf{1}, \nu_t \rangle$, which is zero if $X(t) \geq 1$ and $1 - X(t)$, otherwise. This is captured succinctly by the relation $1 - \langle \mathbf{1}, \nu_t \rangle = [1 - X(t)]^+$.

The input data for the fluid equations includes the arrival rate $\lambda$, and the initial conditions, consisting of the total initial mass in system, and the initial (fluid) age and potential queue measures. Then the space of possible initial conditions for the fluid equations is given by

$$\mathfrak{S} := \left\{ \begin{array}{c} (\tilde{x}, \tilde{\nu}, \tilde{\eta}) \in \mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r) : \\ 1 - \langle \mathbf{1}, \tilde{\nu} \rangle = [1 - \tilde{x}]^+ \end{array} \right\}. \tag{2.5}$$

We now give a precise formulation of the fluid equations introduced in [24] with $E(t) = E^\lambda(t) := \lambda t$ for $t \geq 0$ therein. These equations will also involve the queue process $Q(t)$, which represents the total mass in queue (awaiting service) at time $t$, and the non-decreasing processes $D(t), K(t)$, $S(t)$ and $R(t)$ represent, respectively, the cumulative mass of departures from the queue, entry into service, and respectively, the potentially reneged and actually reneged jobs from the queue in the interval $[0, t]$.

**Definition 2.3. (Fluid Equations)** Given hazard rate functions $h^r$ and $h^s$, the càdlàg function $(X, \nu, \eta)$ 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 with arrival rate $\lambda \geq 0$ and initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$ if for every $t \in [0, \infty)$, we have

$$S(t) := \int_0^t \langle h^r, \eta_s \rangle \, ds < \infty, \qquad D(t) := \int_0^t \langle h^s, \nu_s \rangle \, ds < \infty, \tag{2.6}$$

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

$$\begin{aligned} \langle \varphi(\cdot, t), \nu_t \rangle &= \langle \varphi(\cdot, 0), \nu_0 \rangle + \int_0^t \langle \varphi_s(\cdot, s) + \varphi_x(\cdot, s), \nu_s \rangle \, ds \\ &\quad - \int_0^t \langle h^s(\cdot) \varphi(\cdot, s), \nu_s \rangle \, ds + \int_0^t \varphi(0, s) \, dK(s), \end{aligned} \tag{2.7}$$

where

$$K(t) = \langle \mathbf{1}, \nu_t \rangle - \langle \mathbf{1}, \nu_0 \rangle + D(t); \tag{2.8}$$

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

$$\langle \varphi(\cdot,t),\eta_t\rangle = \langle \varphi(\cdot,0),\eta_0\rangle + \int_0^t \langle \varphi_s(\cdot,s)+\varphi_x(\cdot,s),\eta_s\rangle\, ds - \int_0^t \langle h^r(\cdot)\varphi(\cdot,s),\eta_s\rangle\, ds + \lambda\int_0^t \varphi(0,s)\, ds; \tag{2.9}$$

with the non-idling constraint

$$1-\langle \mathbf{1},\nu_t\rangle = [1-X(t)]^+, \tag{2.10}$$

where

$$X(t) = X(0)+\lambda t - D(t) - R(t), \tag{2.11}$$

*with*

$$R(t) = \int_0^t \left(\int_0^{Q(s)} h^r((F^{\eta_s})^{-1}(y))dy\right) ds, \tag{2.12}$$

where recall $F^{\eta_t}(x) := \eta_t[0,x]$, and $(F^{\eta_t})^{-1}$ denotes the right-continuous inverse defined in (1.3), and

$$Q(t) = X(t) - \langle \mathbf{1},\nu_t\rangle, \tag{2.13}$$

*with $Q$ also satisfying the inequality constraint*

$$Q(t) \le \langle \mathbf{1},\eta_t\rangle. \tag{2.14}$$

**Remark 2.4.** Note that if $(X,\nu,\eta)$ solves the fluid equations with arrival rate $\lambda$, and initial condition $(X(0),\nu_0,\eta_0)\in\mathfrak{S}$, then we also have $(X(t),\nu_t,\eta_t)\in\mathfrak{S}$ for every $t>0$. It is also true that if $\eta_0\in\mathcal{M}_F^c[0,H^r)$, then we also have $\eta_t\in\mathcal{M}_F^c[0,H^r)$ for every $t>0$ (this follows from the expression for $\eta_t$ in (2.18) below, from which it is clear that if $\eta_0$ does not charge points then neither does $\eta_t$).

Also, note from (2.13) and (2.10) that for each $t\in[0,\infty)$,

$$Q(t) = [X(t)-1]^+. \tag{2.15}$$

For future use, we also observe that (2.8), (2.13) and (2.11), when combined, show that for every $t\in[0,\infty)$,

$$Q(0)+\lambda t = Q(t)+K(t)+R(t), \tag{2.16}$$

which is simply a mass conservation equation. In addition, we will find it convenient to define

$$B(t) := \langle \mathbf{1},\nu_t\rangle, \quad t\ge 0, \tag{2.17}$$

which represents the limiting fraction of busy servers.

**Remark 2.5.** Given a solution $(X,\nu,\eta)$, we will refer to $(D,K,R,S,Q,B)$ as auxiliary processes.

We now provide an informal, intuitive explanation for the form of the fluid equations. Note that $\nu_s(dx)$ represents the amount of mass (or fraction of servers) that are processing jobs whose ages lie in the range $[x,x+dx)$ at time $s$, and $h^s(x)$ represents the conditional mean rate at which the mass of jobs with age in $[x,x+dx)$ completes service at time $s$. Hence, in (2.6), $\langle h^s,\nu_s\rangle$ represents the departure rate of mass from the fluid system due to services at time $s$, and its integral, $D(t)$, is the cumulative departure rate due to service completion in the interval $[0,t]$. By an exactly analogous reasoning, the other quantity $S(t)=\int_0^t\langle h^r,\eta_s\rangle ds$ in (2.6) represents the cumulative potential reneging from the system in the interval $[0,t]$. However, the actual reneging rate is restricted to abandonments of those in queue. Since entries into the queue take place in the order of arrival, the age of the oldest (equivalently, head-of-the line) mass in the fluid queue is $\bar{a}_s := (F^{\eta_s})^{-1}(Q(s))$, so that $\eta_s[0,\bar{a}_s]=Q(s)$. Here, recall $F^{\eta_s}$ represents the cumulative distribution function of $\eta_s$. Thus, the actual reneging rate at any time $s$ only counts the mass reneging from the potential queue measure $\eta_s$ whose age lies in the restricted interval $[0,\bar{a}_s]$, rather than the entire interval $[0,\infty)$. A standard change of variable then yields the expression in (2.12). Next, recalling the interpretations of the

quantities $K$, $R$ and $Q$ stated prior to Definition 2.3, note that equations (2.8), (2.13) and (2.11) are simply mass conservation equations, and (2.10) represents a non-idling condition that ensures that no server can idle when there is work in the queue. Moreover, the inequality (2.14) expresses the constraint that at any time $t$, the fluid queue is bounded by the total mass of the fluid potential queue measure, since the latter also includes mass that may have already gone into service (and possibly also departed the system) by that time, provided its patience time exceeds the total time elapsed since arrival. Finally, equations (2.7) and (2.9) govern the evolution of the fluid age measure $\nu$ and potential queue measure $\eta$, respectively. In particular, the second term on the right-hand-side of (2.7) represents the change in $\langle\varphi,\nu\rangle$ over the interval $[0,t]$ due to transport or shift of the ages at unit rate to the right, the third term accounts for changes due to departure of mass from the system due to service, and the last term captures changes due to new entry into system, which are driven by the function $K$, the cumulative entry into service. The equation (2.9) is exactly analogous, but with $h^r$ and the cumulative arrivals $E^\lambda$ into the system in place of $h^s$ and $K$, respectively, and the third term on the right-hand side now representing departure of mass from the system due to potential reneging.

We now state a result that was proved in [23, 25]. Recall the definition of the space $\mathfrak{S}$ given in (2.5).

**Theorem 2.6.** *Suppose Assumption 2.1 holds and fix $\lambda \geq 0$, and $(X(0),\nu_0,\eta_0) \in \mathfrak{S}$. Then there is at most one solution to the fluid equations with arrival rate $\lambda$ and initial condition $(X(0),\nu_0,\eta_0)$, and if $\eta_0 \in \mathcal{M}_F^c[0,H^r)$, then there also exists a continuous solution $(X,\nu,\eta) = \{(X(t),\nu_t,\eta_t), t \geq 0\}$ with arrival rate $\lambda$ and initial condition $(X(0),\nu_0,\eta_0)$. Moreover, given any solution $(X,\nu,\eta) = \{(X(t),\nu_t,\eta_t), t \geq 0\}$ to the fluid equations associated with $\lambda$ and $(X(0),\nu_0,\eta_0) \in \mathfrak{S}$, the following properties hold:*
*(i) for any bounded or nonnegative measurable function $\psi$ on $[0,\infty)$ and for $\psi = h^r$, for every $t \geq 0$,*

$$\int_{[0,H^r)} \psi(x)\eta_t(dx) = \int_{[0,H^r)} \psi(x+t)\frac{\bar{G}^r(x+t)}{\bar{G}^r(x)}\eta_0(dx) + \int_0^t \psi(s)\bar{G}^r(s)\lambda ds; \tag{2.18}$$

*(ii) for any bounded or nonnegative measurable function $\psi$ on $[0,\infty)$ and for $\psi = h^s$, for every $t \geq 0$,*

$$\int_{[0,H^s)} \psi(x)\nu_t(dx) = \int_{[0,H^s)} \psi(x+t)\frac{\bar{G}^s(x+t)}{\bar{G}^s(x)}\nu_0(dx) + \int_0^t \psi(t-s)\bar{G}^s(t-s)dK(s), \tag{2.19}$$

*with $K$ equal to the auxiliary process defined in (2.8) of the fluid equations;*
*(iii) if $Q$ and $B$ are the associated auxiliary processes defined in (2.13) and (2.17) respectively, then $K$ is an absolutely continuous function and for a.e. $t \geq 0$, the derivative $K'$ of $K$ satisfies*

$$K'(t) = k(t) := \begin{cases} \lambda & \text{if } B(t) < 1, \\ \lambda = \langle h^s, \nu_t\rangle & \text{if } B(t) = 1 \text{ and } Q(t) = 0, \\ \langle h^s, \nu_t\rangle & \text{if } B(t) = 1 \text{ and } Q(t) > 0. \end{cases} \tag{2.20}$$

*Proof.* Uniqueness of the solution to the fluid equations follows from Theorem 3.5 of [23] since $(X(0),\nu_0,\eta_0) \in \mathfrak{S}$ implies $(E^\lambda, X(0),\nu_0,\eta_0)$ lies in the space $\mathcal{S}_0$ therein, where recall $E^\lambda(t) = \lambda t$. Likewise, existence of a solution with arrival rate $\lambda$ and initial condition $(X(0),\nu_0,\eta_0) \in \mathfrak{S}$ with $\eta_0 \in \mathcal{M}_F^c[0,H^r)$ can be deduced from Theorem 3.6 of [23], once we justify that the conditions of that theorem are satisfied in the present setting. First, it is not hard to see that for any arrival rate $\lambda \geq 0$ and initial condition $(X(0),\nu_0,\eta_0) \in \mathfrak{S}$ with $\eta_0 \in \mathcal{M}_F^c[0,H^r)$ one can construct a sequence of $N$-server systems with Poisson $(N\lambda)$ arrival process $E^N$ and initial condition $(X^N(0),\nu_0^{(N)},\eta_0^{(N)})$ such that Assumption 3.1 of [23] is satisfied. Second, note that since $\eta_0$ is a continuous measure, and $E^\lambda$ is continuous, Assumption 3.2 of [23] is also satisfied. Finally, Assumption 3.3 of [23] is a direct consequence of Assumption 2.1 of this paper, and thus the application of Theorem 3.6 of [23] is justified.

We now turn to estabishing the properties of any solution $(X,\nu,\eta)$ to the fluid equations. First, note that the forms of both (2.9) and (2.7) are analogous to that of (4.2) in [25], and therefore the integrability conditions in (2.6) imply that (4.1) of [25] holds. Thus, (2.18) and (2.19) for $\psi \in \mathcal{C}_c[0,H^r)$ and $\psi \in \mathcal{C}_c[0,H^s)$, respectively, follow from Theorem 4.1 of [25]. By using a standard approximation argument, namely representing indicators of finite open intervals in $\mathbb{R}_+$ as monotone limits of continuous functions with

compact support and appealing to the monotone class theorem, it follows that both equations in fact hold for any bounded measurable or nonnegative measurable $\psi$. In particular, these equations also hold with $\psi = h^r$ in (2.18) and $\psi = h^s$ in (2.19). The latter fact is used several times in this paper.

We now turn to the proof of property (iii), which is similar in spirit to (3.12) of [25] and Corollary 3.7 of [23]. We supply the details for completeness. First, note that $D, S$ and $R$ are absolutely continuous by definition. Thus, (2.11) shows that $X$ is absolutely continuous, which then implies that $B(t) = \langle \mathbf{1}, \nu_t \rangle = \min(X(t), 1)$ is also absolutely continuous. In turn, by (2.8) and (2.13), this implies that $K$ is absolutely continuous. Further, (2.8), (2.16), (2.6) and (2.17) show that for a.e. $t > 0$,

$$K'(t) = \lambda - Q'(t) - \int_0^{Q(t)} h^r((F^{\eta_t})^{-1}(y))dy, \quad \text{and} \quad K'(t) = B'(t) + \langle h, \nu_t \rangle. \tag{2.21}$$

We now recall the following standard fact that given an absolutely continuous function and a set $A$ on which it is constant, the derivative of the function is zero for almost every $t \in A$. Thus, for almost every $t$ in the set where $B$ is constant, we have $B'(t) = 0$, and (2.21) implies $K'(t) = \langle h, \nu_t \rangle$. On the other hand, for almost every $t$ when $Q(t) = 0$, it follows that $Q'(t) = 0$ and hence, by (2.21) that $K'(t) = \lambda$. Thus, when both $B(t) = 1$ and $Q(t) = 0$, $K'(t) = \lambda = \langle h, \nu_t \rangle$. The remaining claims in (2.20) then follow from the observations that when $B(t) < 1$, one has $Q(t) = 0$ and when $Q(t) > 0$, one has $B(t)$ equal to the constant one, both of which are easily deduced from (2.10) and (2.13). □

We now state a simple result on the action of time-shifts on solutions to the fluid equations. To state the result, which was formulated as Lemma 3.4 of [24], we will need the following notation: for any $t \in [0, \infty)$, define

$$\begin{array}{rclrclrcl} K^{[t]} & := & K(t+\cdot) - K(t), & X^{[t]} & := & X(t+\cdot), & \nu^{[t]} & := & \nu_{t+\cdot}, \\ R^{[t]} & := & R(t+\cdot) - R(t), & \eta^{[t]} & := & \eta_{t+\cdot}, & Q^{[t]} & := & Q(t+\cdot). \end{array}$$

**Lemma 2.7** (Lemma 3.4 of [24])**.** *Suppose Assumption 2.1 holds. Suppose $(X, \nu, \eta) = \{(X(s), \nu_s, \eta_s), s \geq 0\} \in \mathcal{D}_{\mathfrak{S}}[0, \infty)$ solves the fluid equations with arrival rate $\lambda$ and initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$, then for any $t > 0$, $(X^{[t]}, \nu^{[t]}, \eta^{[t]})$ solves the fluid equations with arrival rate $\lambda$ and initial condition $(X(t), \nu_t, \eta_t) \in \mathfrak{S}$, but with $K, R$ and $Q$ replaced with $K^{[t]}, R^{[t]}$ and $Q^{[t]}$, respectively.*

As in [24], we leave the proof to the reader, since it can be verified by just rewriting the fluid equations and invoking the uniqueness result stated in Theorem 2.6.

2.2. **Invariant States and Random Fixed Points of the Fluid Equations.** Let $\nu_*$ and $\eta_*$ be Borel probability measures on $[0, \infty)$ defined as follows:

$$\nu_*[0, x) := \int_0^x \bar{G}^s(y)\, dy, \qquad x \in [0, H^s), \tag{2.22}$$

$$\eta_*[0, x) := \int_0^x \bar{G}^r(y)\, dy, \qquad x \in [0, H^r). \tag{2.23}$$

Note that $\nu_*$ and $\eta_*$ are well defined due to Assumption 2.1. For $\lambda \geq 1$, define the set $\mathcal{X}_\lambda$ as follows:

$$\mathcal{X}_\lambda := \left\{ x \in [1, \infty) : G^r\left( \left(F^{\lambda\eta_*}\right)^{-1}\left((x-1)^+\right)\right) = \frac{\lambda - 1}{\lambda} \right\}, \tag{2.24}$$

and let

$$x_l^\lambda := \inf\{x \in [1, \infty) : \ x \in \mathcal{X}_\lambda\} \qquad \text{and} \qquad x_r^\lambda := \sup\{x \in [1, \infty) : \ x \in \mathcal{X}_\lambda\}.$$

By (2.23), the map $x \to \eta_*[0, x)$ is strictly increasing on $[0, H^r)$, and therefore $(F^{\lambda\eta_*})^{-1}$ is continuous. Since $G^r$ is also continuous, we have $\mathcal{X}_\lambda = [x_l^\lambda, x_r^\lambda]$ is non-empty. Let $\mathcal{I}_\lambda$ be the invariant manifold for the fluid equations, defined by

$$\mathcal{I}_\lambda := \begin{cases} \{(\lambda, \lambda\nu_*, \lambda\eta_*)\} & \text{if } \lambda < 1, \\ \{(x_*, \nu_*, \lambda\eta_*) : x_* \in \mathcal{X}_\lambda\} & \text{if } \lambda \geq 1. \end{cases} \tag{2.25}$$

Our study of the critical and super-critical regimes will be carried out under the following additional assumption on the invariant manifold $\mathcal{I}_\lambda$.

**Assumption 2.8.** *The set $\mathcal{I}_\lambda$ has a single element, which we express as $z_*^\lambda = (x_*^\lambda, (\lambda \wedge 1)\nu_*, \lambda\eta_*)$, where $x_*^\lambda = \lambda$ when $\lambda < 1$ and $x_*^\lambda$ is the unique element of $\mathcal{X}_\lambda$ when $\lambda \geq 1$.*

Note that Assumption 2.8 imposes a non-trivial restriction only when $\lambda \geq 1$. As stated in Lemma 3.1 of [24], a sufficient condition for Assumption 2.8 to hold when $\lambda > 1$ is for the equation $G^r(x) = (\lambda - 1)/\lambda$ to have a unique solution.

Whereas Assumption 2.8 guarantees a unique deterministic fixed point for the fluid equations, to understand the large-time limits of the fluid equations, it turns out to be important to also understand the collection of random fixed points, defined below. We first introduce the notion of a solution to the fluid equations when the input data is random.

**Definition 2.9.** *Given any $\mathfrak{S}$-valued random element $(X(0), \nu_0, \eta_0)$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we say the càdlàg $\mathfrak{S}$-valued stochastic process $(X, \nu, \eta) = \{(X(t), \nu_t, \eta_t), t \geq 0\}$ is a solution to the fluid equations with arrival rate $\lambda$ and random initial condition $(X(0), \nu_0, \eta_0)$ if for each $\omega \in \Omega$, the function $(X(\omega), \nu(\omega), \eta(\omega)) = \{(X(t, \omega), \nu_t(\omega), \eta_t(\omega)), t \geq 0\}$ solves the fluid equations with arrival rate $\lambda$ and initial condition $(X(0, \omega), \nu_0(\omega), \eta_0(\omega))$.*

**Definition 2.10.** *For $\lambda > 0$, a probability measure $\mu$ on $\mathfrak{S}$ is said to be a random fixed point of the fluid equations with arrival rate $\lambda$ if given any $\mathfrak{S}$-valued random element $(\tilde{X}, \tilde{\nu}, \tilde{\eta})$ whose law is $\mu$, there exists a solution $(X, \nu, \eta)$ to the fluid equations with arrival rate $\lambda$ and random initial condition $(\tilde{X}, \tilde{\nu}, \tilde{\eta})$ such that for each $t \geq 0$, the law of $(X(t), \nu_t, \eta_t)$ is equal to $\mu$.*

**Remark 2.11.** Under our assumptions, a random fixed point always exists. Indeed, it follows from Theorem 5.5 of [24] that the set $\mathcal{I}_\lambda$ in (2.25) describes the so-called invariant manifold (or collection of deterministic fixed points) of the fluid equations. Since for $\lambda \geq 0$, $\mathcal{X}_\lambda$ is always non-empty, an immediate consequence is that for any $z \in \mathcal{I}_\lambda$, the measure $\delta_z$ is a random fixed point of the fluid equations with arrival rate $\lambda$. Moreover, under Assumption 2.8, $\delta_{z_*^\lambda}$ is the only random fixed point that is degenerate (i.e., which concentrates all its mass on one point). A key question we address in this article is to determine conditions under which this is in fact the only random fixed point with arrival rate $\lambda$. As shown in Proposition 4.16 below, a sufficient condition for this to hold is that any solution $(X, \nu, \eta)$ to the fluid equations with arrival rate $\lambda$ and initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$ and auxiliary process $B = \langle \mathbf{1}, \nu \rangle$ satisfies $\eta_t \Rightarrow \lambda\eta_*$ and $B_t \to \lambda \wedge 1$, as $t \to \infty$.

## 3. Assumptions and Main Results

We now state our main results, which require the following additional condition on the service distribution.

**Assumption 3.1.** *The cumulative distribution function $G^s$ of the service distribution has a density $g^s$ and the hazard rate function $h^s = g^s/\bar{G}^s$ satisfies one of the following:*

1. *The quantities $\varepsilon_h := \operatorname{ess\,inf}_{x \geq 0} h^s(x) > 0$ and $c_h := \operatorname{ess\,sup}_{x \geq 0} h^s(x) < \infty$.*
2. *The function $h^s$ is decreasing.*

The second part of the above assumption should be understood in the sense of Remark 2.2, namely $h^s$ is a.e. equal to a decreasing function from $[0, H^s)$ to $\mathbb{R}_+$. Note that under both parts of the assumption, the hazard rate function $h^s$ has a finite essential supremum. Since the hazard rate function of any distribution is only locally integrable and never integrable on its support, both Assumptions 3.1(1) and 3.1(2) imply $H^s = \infty$.

**Theorem 3.2.** *Suppose Assumption 2.1 holds, and $\nu_*$ and $\eta_*$ are as defined in (2.22) and (2.23). Also, suppose $(X, \nu, \eta)$ solves the fluid equations with arrival rate $\lambda$ and initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$, with auxiliary processes $(D, K, R, S, Q, B)$ as in Remark 2.5. Then the following is true:*

(1) *When $\lambda < 1$, it follows that $(X_t, \nu_t, \eta_t) \to (\lambda, \lambda\nu_*, \lambda\eta_*)$ as $t \to \infty$. In particular, $\delta_{z_*^\lambda}$, with $z_*^\lambda = (\lambda, \lambda\nu_*, \lambda\eta_*)$, is the unique random fixed point of the fluid equations with arrival rate $\lambda$.*

(2) *When $\lambda > 1$, and Assumption 3.1 is also satisfied, then $\eta_t \Rightarrow \lambda\eta_*$ as $t \to \infty$ and*

(a) *there exists $T < \infty$ such that*

$$B(t) = \langle \mathbf{1}, \nu_t \rangle = 1, \quad \text{for all } t \geq T, \tag{3.1}$$

*and*

$$\nu_t \Rightarrow \nu_* \text{ and } \langle h^s, \nu_t \rangle \to 1 \quad \text{as } t \to \infty, \tag{3.2}$$

*with the convergence in (3.2) also holding in total variation when Assumption 3.1(1) holds;*

(b) *if, in addition, Assumption 2.8 is also satisfied (with $z_*^\lambda$ as defined therein), then $\delta_{z_*^\lambda}$ is the unique random fixed point of the fluid equations with arrival rate $\lambda$.*

(3) *If $\lambda = 1$ and Assumption 3.1(2) is satisfied, then $\eta_t \Rightarrow \lambda\eta_*$ and $B(t) \to 1$ as $t \to \infty$. If, in addition, Assumption 2.8 holds (with $z_*^1$ as defined therein), then $\delta_{z_*^1}$ is the unique random fixed point of the fluid equations with arrival rate 1.*

*Proof.* The proof of statement (1) is given in Section 4.1, and for all $\lambda \geq 0$, the (weak) convergence of $\eta_t$ to $\lambda\eta_*$ under Assumption 3.1 follows from Lemma 4.1. Now, when $\lambda > 1$, (3.1) and (3.2) follow from Proposition 4.8 and Remark 4.7 when Assumption 3.1(1) holds, and from Proposition 4.12 and Lemma 4.15 when Assumption 3.1(2) holds. Further, when $\lambda = 1$ the convergence $B(t) \to 1$ under Assumption 3.1(2) also follows from Proposition 4.12. Lastly, the uniqueness results for the random fixed point stated in (2b) and (3) follow from the convergence results in (2a) and (3) and Proposition 4.16.

□

**Remark 3.3.** The main application of Theorem 3.2 is to characterize the limit of the scaled stationary distributions of the sequence of $N$-server measure-valued state processes, and thereby (partially) fix a technical flaw in the convergence result stated in Theorem 3.3 of [24].

To explain this in greater detail, let $\bar{Z}_*^{(N)} := (\bar{X}_*^{(N)}, \bar{\nu}_*^{(N)}, \bar{\eta}_*^{(N)})$ have the law of the stationary distribution of the measure-valued $N$-server state dynamics of an $N$-server queue with reneging introduced in [23], when the scaled arrival process is given by $\bar{E}^{(N)}$ (e.g., Poisson with a scaled arrival rate $\lambda^{(N)} > 0$). Existence of such a stationary distribution was established in Theorem 7.1 of [24]. Also, let $\bar{Z}^{(N)} := (\bar{X}^{(N)}, \bar{\nu}^{(N)}, \bar{\eta}^{(N)})$ represent the dynamics of the fluid-scaled measure-valued state representation of the $N$-server queue with initial condition $\bar{Z}^{(N)}(0) = \bar{Z}_*^{(N)}$. Then, under the assumption that $\bar{E}^{(N)}$ converges weakly to $E^\lambda$ for some $\lambda > 0$, tightness of the sequence $\{\bar{Z}_*^{(N)} := (\bar{X}_*^{(N)}, \bar{\nu}_*^{(N)}, \bar{\eta}_*^{(N)})\}_{N\in\mathbb{N}}$, was established in Theorem 6.2 of [24]. Let $\bar{Z} = (\bar{X}, \bar{\nu}, \bar{\eta})$ denote any subsequential limit. We now claim that then (the law of) $\bar{Z}$ must be a random fixed point of the fluid equations with arrival rate $\lambda$. To see why the claim is true, we invoke the fluid limit theorem established in Theorem 3.6 of [23], to conclude that for any $t > 0$, the $N$-server fluid-scaled state process $\bar{Z}^{(N)}(t)$ (initialized at the stationary distribution $\bar{Z}_*^{(N)}$) converges weakly to $Z(t)$, where $Z := (X, \nu, \eta)$ solves the fluid equations with arrival rate $\lambda$ and initial condition $\bar{Z}$. However, for any $t > 0$, since by stationarity $\bar{Z}^{(N)}(t)$ has the same law as $\bar{Z}_*^{(N)}$, it follows that the laws of their corresponding weak limits, $Z(t)$ and $\bar{Z}$, must also coincide. By Definition 2.10, this proves the claim that $\bar{Z}$ is a random fixed point.

In the proof of Theorem 3.3 (of Section 6.2) in [24], it was assumed without justification that $\bar{Z}$ is deterministic, and that was used to conclude that $\bar{Z}$ must belong to the invariant manifold $\mathcal{I}_\lambda$ (see Remark 2.11). When combined with Assumption 2.8, this leads to the conclusion that $\bar{Z} = z_*^\lambda$, thus showing that all subsequential limits coincide, and hence, that $z_*^\lambda$ is the weak limit of the original stationary sequence $(\bar{Z}_*^{(N)})_{N\in\mathbb{N}}$. However, one cannot assume *a priori* that $\tilde{Z}$ is deterministic, and, as argued above, one only knows that any subsequential limit is a *random fixed point.* To make this argument complete, which was one of the main motivations of this paper, one needs to show that there is precisely one random fixed point, namely the one concentrated at $z_*^\lambda$. Theorem 3.2 does precisely this for the class of service distributions

satisfying Assumption 3.1, thus closing the gap in the proof of the convergence result in [24] (for service distributions in that class). However, this still leaves the open question of whether this result remains true for a larger class of service distributions, in particular the entire class considered in [24].

**Remark 3.4.** Further, a related ancillary goal of this work is to determine whether the diagram in Figure 1 below commutes under general convergence conditions on the initial states (essentially Assumption 3.1 of [23]). Referring to the same notation as used in Remark 3.3, the top horizontal arrow in Figure 1 holds due to ergodicity of the $N$-server state dynamics, which was established in Theorem 7.1 of [24] under some additional conditions on the service and reneging distributions (see Assumption 7.1 therein). On the other hand, as already mentioned in Remark 3.3, the left vertical arrow follows from the fluid limit theorem Theorem 3.6 of [23] (under suitable convergence assumptions on the initial data).

$$\begin{array}{ccc}
\bar{Z}^{(N)}(t) = (\bar{X}^{(N)}(t), \bar{\nu}_t^{(N)}, \bar{\eta}_t^{(N)}) & \overset{\textit{Thm 7.1 of [24]}}{\Longrightarrow} & \bar{Z}_*^{(N)} = (\bar{X}_*^{(N)}, \bar{\nu}_*^{(N)}, \bar{\eta}_*^{(N)}) \\
\Big\| & & \Big\| \\
\underset{\textit{Thm 3.6 of [23]}}{\Downarrow} & & \underset{\textit{Thm 6.2 of [24] and } \textbf{Thm 3.2}}{\Downarrow} \\
Z(t) = (X(t), \nu_t, \eta_t) & \overset{\textbf{Thm 3.2}}{\longrightarrow} & z_* = (x_*, (\lambda \wedge 1)\nu_*, \lambda\eta_*)
\end{array}$$

FIGURE 1. Interchange of Limits Diagram

Along with the tightness of $(\bar{Z}_*^{(N)})_{N\in\mathbb{N}}$ established in [24], Theorem 3.2 of the present article completes the diagram by establishing (for a class of service distributions) the right vertical arrow (as explained in Remark 3.3) as well as the bottom horizontal arrow, though the latter only when $\lambda \neq 1$ (i.e., in the subcritical and supercritical regimes). It would be worthwhile in the future to investigate whether this result can be extended further, in particular to establish converence even in the critical regime $\lambda = 1$, possibly under additional conditions such as a finite second moment condition, like that imposed in Theorem 3.9 of [25] (to study large-time behavior of fluid limits in the absence of reneging).

## 4. Proof of Theorem 3.2

We assume throughout this section that Assumption 2.1 holds. We then have the following elementary lemma.

**Lemma 4.1.** *Fix $\lambda \geq 0$ and, given any $\eta_0 \in \mathcal{M}_F[0, H^r)$, let $\eta = (\eta_t)_{t\geq 0}$ be the solution to* (2.9)*. Then $\eta_t \Rightarrow \lambda\eta_*$ as $t \to \infty$.*

*Proof.* Fix $\psi \in \mathcal{C}_b(\mathbb{R}_+)$. In view of (2.18), the boundedness of $\psi$, the finiteness of the measure $\eta_0$ the dominated convergence theorem and the fact that $\bar{G}^r(x+t)/\bar{G}^r(x) \to 0$ for every $x \in [0, H^r)$ as $t \to \infty$, together imply that the first term on the right-hand side of (2.18) vanishes. On the other hand, since the mean patience time $\int_0^\infty \bar{G}^r(s)ds$ is finite, the dominated convergence theorem shows that the last term on the right-hand side of (2.18) converges to $\langle \psi, \lambda\eta_* \rangle$. This concludes the proof that $\eta_t \Rightarrow \lambda\eta_*$ as $t \to \infty$. □

4.1. **Proof in the Subcritical Regime.** In this section we prove part (1) of Theorem 3.2. Fix $\lambda \in (0, 1)$ and $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$. Suppose $(X, \nu, \eta)$ is a solution to the fluid equations, and let $(D, K, R, S, Q, B)$ be the corresponding auxiliary processes.

The weak convergence of $\eta_t$ to $\lambda\eta_*$ as $t \to \infty$ follows from Lemma 4.1. We now analyze the remaining components of the solution. Using the definition of $D$ from (2.6), setting $\psi = h^s$ in (2.19), interchanging the

order of integration, using integration by parts and the fact $G^s(0+) = 0$, we obtain

$$
\begin{aligned}
D(t) = \int_0^t \langle h^s, \nu_s \rangle \, ds &= \int_0^t \left( \int_{[0,H^s)} \frac{g^s(x+s)}{\bar{G}^s(x)} \nu_0(dx) + \int_{[0,s]} g^s(s-u) dK(u) \right) ds \\
&= \int_{[0,H^s)} \frac{G^s(t+x) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) + \int_{[0,t]} G^s(t-u) dK(u) \\
&= \int_{[0,H^s)} \frac{G^s(t+x) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) + \int_0^t K(s) g^s(t-s) ds.
\end{aligned} \tag{4.1}
$$

Substituting this in (2.11), using (2.16) and (2.12), and performing repeated integration by parts, we obtain

$$
\begin{aligned}
X(t) &= X(0) + \lambda t - \int_{[0,H^s)} \frac{G^s(t+x) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) - \int_0^t K(s) g^s(t-s) ds - R(t) \\
&= X(0) + \lambda t - \int_{[0,H^s)} \frac{G^s(t+x) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) \\
&\qquad - \int_0^t (Q(0) + \lambda s - Q(s) - R(s)) g^s(t-s) ds - R(t) \\
&= X(0) - Q(0) G^s(t) - \int_{[0,H^s)} \frac{G^s(t+x) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) + \int_0^t Q(s) g^s(t-s) ds \\
&\qquad + \lambda \int_0^t \bar{G}^s(t-s) ds + \int_0^t R(s) g^s(t-s) ds - R(t) \\
&= X(0) - Q(0) G^s(t) - \int_{[0,H^s)} \frac{G^s(t+x) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) + \int_0^t Q(s) g^s(t-s) ds \\
&\qquad + \int_0^t \left( \lambda - \int_0^{Q(s)} h^r((F^{\eta_s})^{-1}(y)) dy \right) \bar{G}^s(t-s) ds,
\end{aligned} \tag{4.2}
$$

which implies that for each $t \geq 0$,

$$
X(t) \leq X(0) - Q(0) G^s(t) - \int_{[0,H^s)} \frac{G^s(t+x) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) + \lambda \int_0^t \bar{G}^s(u) du + \int_0^t Q(s) g^s(t-s) ds. \tag{4.3}
$$

We now make use of the following simple observation.

**Lemma 4.2.** $\limsup_{t\to\infty} \int_0^t Q(s) g^s(t-s) ds \leq \limsup_{t\to\infty} Q(t)$.

*Proof.* Let $q := \limsup_{t\to\infty} Q(t)$. Then for each $\epsilon > 0$, there exists $T_\epsilon < \infty$ such that $Q(t) \leq q + \epsilon$ for all $t \geq T_\epsilon$. So for each $t > T_\epsilon$, it follows that

$$
\begin{aligned}
\int_0^t Q(s) g^s(t-s) ds &= \int_0^{T_\epsilon} Q(s) g^s(t-s) ds + \int_{T_\epsilon}^t Q(s) g^s(t-s) ds \\
&\leq \left( \sup_{0 \leq s \leq T_\epsilon} Q(s) \right) (G^s(t) - G^s(t - T_\epsilon)) + (q + \epsilon) G^s(t - T_\epsilon).
\end{aligned}
$$

By taking the limit supremum as $t \to \infty$ of both sides, we have $\limsup_{t\to\infty} \int_0^t Q(s) g^s(t-s) ds \leq q + \epsilon$. The lemma follows on taking $\epsilon \to 0$. □

Continuing with the proof of Theorem 3.2(1), taking the limit supremum in (4.3), and using Lemma 4.2, the identity $\int_0^\infty G^s(u) du = 1$ from Assumption 2.1, the fact that $\lim_{t\to\infty} (G^s(t+x) - G^s(x))/\bar{G}^s(x) \to 1$ for every $x$, the bounded convergence theorem and the identity $X(0) = Q(0) + \langle \mathbf{1}, \nu_0 \rangle$ from (2.13), we obtain

$$
\limsup_{t\to\infty} X(t) \leq \lambda + \limsup_{t\to\infty} \int_0^t Q(s) g^s(t-s) ds \leq \lambda + \limsup_{t\to\infty} Q(t). \tag{4.4}
$$

We now claim that there exists $T' < \infty$ such that $\langle 1, \nu_t \rangle < 1$ for all $t \geq T'$. We argue by contradiction to prove the claim. If the claim is false, note that for any $T' < \infty$, there would exist $T > T'$ such that $\langle 1, \nu_T \rangle = 1$. Then, due to (2.15), we would have $\limsup_{t\to\infty} X(t) = \limsup_{t\to\infty} Q(t) + 1$, which contradicts (4.4) since $\lambda < 1$. Thus, fix $T' < \infty$ as in the claim. Then, by Lemma 2.7, $(X^{[T']}, \nu^{[T']}, \eta^{[T']})$ solves the fluid equations with arrival rate $\lambda$ and initial condition $(X(T'), \nu_{T'}, \eta_{T'})$ and hence, (2.19) holds with $\nu$ and $K$ replaced with $\nu^{[T']}$ and $K^{[T']}$, respectively. Since $\nu^{T'}(t) = \nu_{T'+t}$ and by (2.15), (2.12) and (2.16), $Q(T' + \cdot) \equiv 0$, $R^{[T']}(\cdot) \equiv 0$ and $K^{T'}(t) = K(T' + t) - K(T') = \lambda t$, $t \geq 0$, this implies that for every $\psi \in \mathcal{C}_b[0, H^s)$,

$$\int_{[0,H^s)} \psi(x)\nu_{T'+t}(dx) = \int_{[0,H^s)} \psi(x+t)\frac{\bar{G}^s(x+t)}{\bar{G}^s(x)}\nu_{T'}(dx) + \int_0^t \psi(t-s)\bar{G}^s(t-s)\lambda ds.$$

Then, arguing as in the proof of Lemma 4.1, sending $t \to \infty$, and invoking the bounded convergence theorem, the first integral on the right-hand side vanishes, and the second integral converges to $\lambda \int_{[0,H^s)} \psi(x)\bar{G}^s(x)dx$. Recalling that $\nu_*(dx) = \bar{G}^s(x)dx$ and $\int_0^\infty \bar{G}^s ds = 1$ from Assumption 2.1, it follows that $\nu_t \Rightarrow \lambda\nu_*$. In turn, by the continuous mapping theorem this implies $\langle \mathbf{1}, \nu_t \rangle \Rightarrow \lambda$ as $t \to \infty$. When combined with (2.10) and the fact that $\lambda < 1$, this implies that as $t \to \infty$, the weak limits of $X(t)$ and $\langle \mathbf{1}, \nu_t \rangle$ coincide and are equal to $\lambda$. This concludes the proof of the first assertion of Theorem 3.2(1).

Now, if the initial condition $(X(0), \eta_0, \nu_0)$ had the law $\mu$ of a random fixed point with arrival rate $\lambda < 1$, then the convergence just established would imply that $\mathbb{P}(\eta_0 = \lambda\eta_*) = 1$ and $\mathbb{P}(\nu_0 = \lambda\nu_*) = 1$. By the continuous mapping theorem, the latter implies that almost surely $\langle \mathbf{1}, \nu_0 \rangle = \langle \mathbf{1}, \lambda\nu_* \rangle = \lambda$. Since $\lambda < 1$, it then follows from (2.10) that $X(0) = \lambda$ almost surely, thus proving that $\mu = \delta_{z_*^\lambda}$ with $z_*^\lambda = (\lambda, \lambda\eta_*, \lambda\nu_*)$. This completes the proof of Theorem 3.2(1).

4.2. **Proof of Theorem 3.2(2) when the hazard rate function is bounded away from zero and infinity.** In this section we prove Theorem 3.2(2a) under Assumption 3.1(1). Fix $\lambda > 1$, suppose Assumption 2.1 is satisfied, Assumption 2.8 holds (with $z_*^\lambda = (x_*^\lambda, \nu_*, \lambda\nu_*)$ denoting the unique element of $\mathcal{I}_\lambda$) and Assumption 3.1(1) holds (with associated positive constants $\varepsilon_h > 0$, $c_h < \infty$). For notational convenience, we shall denote by $f^*(x) = \bar{G}^s(x)$ the density of $\nu^*$. Note that the lower bound on $h^s$ implies that $g^s$, and thus $f^* = \bar{G}^s$, is strictly positive on $(0, \infty)$.

Now, fix the initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$, and suppose $(X, \nu, \eta)$ is the associated solution to the fluid equations. We will establish convergence, as $t \to \infty$, of the fluid age measure $\nu_t$ described by (2.7) using an extended relative entropy functional in a manner reminiscent of a Lyapunov function. Recall that $\mathcal{P}(E)$ denotes the space of probability measures on a measurable space $E$, and for a *finite* measure $P$ on $E$, define the functional $R : (P\|\cdot) : \mathcal{P}(E) \mapsto (-\infty, \infty]$ by

$$R(P\|Q) := \begin{cases} \displaystyle\int_E \log\frac{dP}{dQ}(x)dP(x) & \text{if } P \ll Q, \\ \infty & \text{otherwise,} \end{cases} \tag{4.5}$$

where $P \ll Q$ means $P$ is absolutely continuous with respect to $Q$ and we use the convention $0 \log 0 = 0$. We emphasize that we do not require $P$ to be a probability measure, as we will often have to deal with sub-probability measures, but when both $P$ and $Q$ are probability measures, this is simply the relative entropy functional.

**Remark 4.3.** If $c_P = P(E) > 0$ denotes the total mass of $P$, then writing the above integral as $\int_E \frac{dP}{dQ} \log \frac{dP}{dQ} dQ$ and using the convexity of $x \mapsto x \log x$ on $(0, \infty)$ gives the lower bound

$$R(P\|Q) \geq c_P \log c_P, \tag{4.6}$$

which is attained by $P$ that is a constant multiple of the probability measure $Q$. In particular, $R(P\|Q)$ may assume negative values. However, when $P$ is a probability measure, $R(P\|Q)$ is always nonnegative and $R(P\|Q) = 0$ holds if and only if $P = Q$.

The proof of Theorem 3.2(2) will make use of the following properties of the extended relative entropy functional.

**Lemma 4.4.** *Suppose $P$ and $Q$ are finite measures on $\mathbb{R}_+$, equipped with the Borel $\sigma$-algebra, with $c_P := P(\mathbb{R}_+) > 0$ and $Q(\mathbb{R}_+) = 1$. If $P$ and $Q$, respectively, have densities $p$ and $q$ (with respect to Lebesgue measure), then*

$$\int_0^\infty |p(x) - q(x)|dx \leq |c_P - 1| + \Big(2c_P^{-1}|R(P\|Q)| + 2|\log c_P|\Big)^{1/2}. \tag{4.7}$$

*Proof.* First note that $c_P^{-1}P$ and $Q$ are probability measures, and so, invoking Pinsker's inequality (see, e.g., [15], p. 44) in the second inequality below, we obtain

$$\begin{aligned}\int_0^\infty |p(x) - q(x)|dx &\leq \int_0^\infty |p(x) - c_P^{-1}p(x)|dx + \int_0^\infty |c_P^{-1}p(x) - q(x)|dx \\ &\leq |c_P - 1| + \big(2R(c_P^{-1}P\|Q)\big)^{1/2} \\ &\leq |c_P - 1| + \big(2c_P^{-1}R(P\|Q) - 2\log c_P\big)^{1/2}\end{aligned}$$

which is clearly dominated by the right-hand side of (4.7). □

The second property is encapsulated in the following lemma, which crucially relies on the lower bound on the hazard rate $h^s$, and whose proof is relegated to Appendix A.

**Lemma 4.5.** *Let $f : [0,\infty) \mapsto [0,\infty)$ be a measurable function that satisfies $\int_0^\infty f dx \leq 1$, suppose $z_f := \int_0^\infty h^s f dx < \infty$ and $\mu^f$ is the measure with density $f$. Then*

$$\int_0^\infty h^s(x)f(x)\log\frac{f(x)}{f^*(x)}dx - z_f\log z_f \geq \varepsilon_h\int_0^\infty f(x)\log\frac{f(x)}{f^*(x)}dx = \varepsilon_h R(\mu^f\|\nu_*). \tag{4.8}$$

The proof of Theorem 3.2(2) is somewhat involved and given in Section 4.2.2. To help make some of those calculations more transparent, first in Section 4.2.1 we carry out some formal calculations (under more stringent conditions) to provide intuition into why the extended relative entropy functional $R(\cdot\|\nu^*)$ may be a good candidate Lyapunov function for the problem at hand (see also Remark 4.6).

4.2.1. *A Formal Calculation.* Observe that equation (2.7) characterizes $(\nu_t)_{t\geq 0}$ as a weak solution to a transport equation. Now, for the purposes of this formal calculation only, suppose that $\nu_0$ has a density, denoted by $f_0$, and for each $t > 0$, suppose the measure $\nu_t$ has a sufficiently smooth density, denoted by $f(x,t)$, $x \geq 0$. For conciseness, below we will use $f(\cdot,t)$ to denote the function $x \mapsto f(x,t)$. Then by (2.17) and (2.6), $\langle \mathbf{1}, \nu_t\rangle = \int_0^\infty f(\cdot,t)dx$ and $\langle h, \nu_t\rangle = \int_0^\infty h^s f(\cdot,t)dx$, the transport equation could be formally rewritten as the following partial differential equation (PDE):

$$\partial_t f(x,t) = -\partial_x f(x,t) - h^s(x)f(x,t), \quad x > 0, t > 0, \tag{4.9}$$

with the boundary condition $f(0,t) = K'(t)$, which by (2.20), takes the form

$$f(0,t) = \begin{cases}\lambda & \text{if } \displaystyle\int_0^\infty f(\cdot,t)dx < 1, \\ \displaystyle\int_0^\infty h^s f(\cdot,t)dx & \text{if } \displaystyle\int_0^\infty f(\cdot,t)dx = 1,\end{cases} \tag{4.10}$$

and the initial condition

$$f(x,0) = f_0(x), \quad x > 0. \tag{4.11}$$

Proceeding with purely formal calculations to gain intuition, note that $f^* = e^{-J}$, where $J(x) := \int_0^x h^s(y)dy < \infty$ for every $x > 0$. For $t > 0$, define

$$r_t := R(\nu_t \| \nu^*) = \int_0^\infty f(\cdot,t) \log \frac{f(\cdot,t)}{f^*} dx = \int_0^\infty f(\cdot,t)(\log f(\cdot,t) + J)dx.$$

Taking derivatives of both sides of the last equation with respect to $t$, and using (4.9), we see that

$$\begin{aligned}\frac{d}{dt} r_t &= \int_0^\infty \partial_t f(\cdot,t)(\log f(\cdot,t) + J + 1)dx \\ &= -\int_0^\infty (\partial_x f(\cdot,t) + h^s f(\cdot,t))(\log f(\cdot,t) + J + 1)dx.\end{aligned}$$

Since $f(\cdot,t)$ is integrable and $H^s = \infty$, it follows that $\liminf_{x\to\infty} f(x,t) = 0$. Using integration by parts, and assuming (without justification) that $\lim_{x\to\infty} f(x,t)(\log f(x,t) + J(x)) = 0$, we conclude that

$$\begin{aligned}\int_0^\infty (\partial_x f(\cdot,t)(\log f(\cdot,t) + J + 1)dx &= -f(0,t)(\log f(0,t) + 1) - \int_0^\infty f(\cdot,t)\left(\frac{\partial_x f(\cdot,t)}{f(\cdot,t)} + h^s\right)dx \\ &= -f(0,t)\log f(0,t) - \int_0^\infty h^s f(\cdot,t)dx,\end{aligned}$$

On combining the last two equations, and recalling that $J = -\log f^*$, we obtain

$$\begin{aligned}\frac{d}{dt} r_t &= f(0,t)\log f(0,t) - \int_0^\infty h^s f(\cdot,t)(\log f(\cdot,t) + J)dx \\ &= f(0,t)\log f(0,t) - \int_0^\infty h^s f(\cdot,t)\log\frac{f(\cdot,t)}{f^*}dx.\end{aligned}$$

Since $\int_0^\infty f(\cdot,t)dx = \langle \mathbf{1}, \nu_t\rangle \le 1$ and for almost every $t \in [0,\infty)$, (2.6) implies that $\int_0^\infty hf(\cdot,t)dx < \infty$ for such $t$. we can apply the estimate (4.8) from Lemma 4.5 with $f = f(\cdot,t)$ to obtain

$$\int_0^\infty h^s f(\cdot,t)\log\frac{f(\cdot,t)}{f^*}dx \ge \left(\int_0^\infty h^s f(\cdot,t)dx\right)\log\left(\int_0^\infty h^s f(\cdot,t)dx\right) + \varepsilon_h \int_0^\infty f(\cdot,t)\log\frac{f(\cdot,t)}{f^*}dx.$$

Substituting this into the previous display and using the boundary condition (4.10), we have

$$\frac{d}{dt} r_t \le \begin{cases} -\varepsilon_h r_t + \lambda\log\lambda - \left(\int_0^\infty h^s f(\cdot,t)dx\right)\log\left(\int_0^\infty h^s f(\cdot,t)dx\right) & \text{if } \int_0^\infty f(\cdot,t)dx < 1, \\ -\varepsilon_h r_t & \text{if } \int_0^\infty f(\cdot,t)dx = 1. \end{cases} \tag{4.12}$$

This estimate does not directly imply the convergence of $r_t$ to zero. However, the fact that it takes the form $\frac{dr_t}{dt} \le -\varepsilon_h r_t$ in the case $\int_0^\infty f(\cdot,t)dx = 1$ is a sign that the approach might be useful, especially in the supercritical case ($\lambda > 1$), where one might expect that for sufficiently large $t$, $\int_0^\infty f(\cdot,t)dx = 1$. However, translating this intuition into a proof is not straightforward. The rigorous argument provided in the next section indeed derives a version of (4.12) (with some extra error terms), and copes with the more complicated structure of the estimate in the case $\int_0^\infty f(\cdot,t)dx < 1$, as well as the fact that $r_t$ can go negative.

**Remark 4.6.** Note that the PDE (4.9)-(4.10) has some similarities with the age-structured model in equation (3) of [32], with $\nu = 0$ and $d = b = h^s$, except that the boundary condition (4.10) is more complicated. In particular, it is discontinuous due to the appearance of the term $\lambda$ when $\int_0^\infty f(\cdot,t)dx < 1$. Furthermore, although $f^*$ can indeed be seen as an eigenfunction corresponding to the eigenvalue 0 of the stationary equation (which corresponds to equation (7) of [32], again with $\nu = 0$ and $d = b = h^s$), since the hazard rate function $h^s$ is never integrable on $[0,\infty)$, the solution to the dual equation (see (8) of [32]) appears not to be well-defined. Thus, the results of [32] are not applicable to this setting. Furthermore, a rigorous proof cannot in any case rely on an analysis of the PDE because for general initial condition $\nu_0 \in \mathcal{M}_F[0,\infty)$, the measures $\nu_t, t > 0$, need not have densities, and even when they do, their densities have discontinuities in both variables (these discontinuities will be apparent in the rigorous proof in the next section). Nevertheless, along with the calculations given above, this loose analogy further suggests that the extended relative entropy functional may still serve as a Lyapunov function for the dynamics. That verification of this property is

non-trivial will be apparent on noting that it requires additional conditions on $h^s$ and also a restriction to the supercritical regime $\lambda > 1$. In particular, it would be interesting to see if the argument presented in the next section, or a modification thereof, could relax conditions on $h^s$ to address a larger class of service distributions, and also address the critical regime $\lambda = 1$, which currently we only address when the hazard rate function is decreasing (see Proposition 4.12).

4.2.2. *Proof of Theorem 3.2(2).* Fix $\lambda > 1$ and recall the initial condition and associated solution $(X, \nu, \eta)$ to the fluid equations. We start with the proof of part (a). First, note that the limit $\eta_t \Rightarrow \lambda\eta_*$ in (3.1) follows from Lemma 4.1. To establish the remaining limits, we begin with the representation for the age measure $\nu_t$ given in (2.19), which shows that $\nu_t = \theta_t + \mu_t$, where $\theta_t, \mu_t \in \mathcal{M}_F[0,\infty)$, are defined by

$$\langle \psi, \theta_t \rangle := \int_{[0,\infty)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \psi(x+t)\nu_0(dx) \quad \text{and} \quad \langle \psi, \mu_t \rangle := \int_0^\infty \psi(x)\tilde{f}(x,t)dx, \tag{4.13}$$

for every $\psi \in \mathcal{C}_b[0,\infty)$ and $\psi = h^s$, where for all $t \geq 0$,

$$\tilde{f}(x,t) := \begin{cases} \bar{G}^s(x)k_{t-x} & x \in [0,t], \\ 0 & x \in (t,\infty), \end{cases} \tag{4.14}$$

where we recall that $k$, defined in (2.20), is a.e. equal to the derivative $K'$ of $K$.

Now, to estimate $d_{\mathrm{TV}}(\mu_t, \nu_*)$, recall that $f^* = \bar{G}^s$ is the density of $\nu^*$, and so both $\mu_t$ and $\nu_*$ are absolutely continuous with respect to Lebesgue measure. Thus, Lemma 4.4 shows that

$$d_{\mathrm{TV}}(\mu_t, \nu_*) = \int_0^\infty \left| \tilde{f}(x,t) - f^*(x) \right| dx \leq |\langle \mathbf{1}, \mu_t \rangle - 1| + \left(2\langle \mathbf{1}, \mu_t \rangle^{-1}|r_t| + 2|\log\langle \mathbf{1}, \mu_t \rangle|\right)^{1/2}, \tag{4.15}$$

where, for $t \geq 0$,

$$r_t := R(\mu_t \| \nu^*) = \int_0^\infty \tilde{f}(x,t) \log \frac{\tilde{f}(x,t)}{f^*(x)} dx \tag{4.16}$$

$$= \int_0^t \bar{G}^s(t-x)k_x \log k_x dx, \tag{4.17}$$

with the last equality using the fact that $k_{t-x} = \tilde{f}(x,t)/f^*(x)$ due to (4.14). Since the expression in (2.20) and Assumption 3.1(1) show that $k$ is strictly positive and bounded above by $\lambda \vee c_h = \lambda \vee \sup_{x\in[0,\infty)} h(x)$, $r_t$ is well defined and finite.

**Remark 4.7.** Due to the pointwise convergence $\frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \to 0$ as $t \to \infty$, the dominated convergence theorem shows that $\langle \mathbf{1}, \theta_t \rangle$, the total mass of $\theta_t$, converges to zero as $t \to \infty$. Hence, $\theta_t$ converges to the zero measure in total variation. Together with (4.15), it follows that in order to show $B_t = \langle \mathbf{1}, \nu_t \rangle \to 1$ and $d_{\mathrm{TV}}(\nu_t, \nu_*) \to 0$ (and hence, $\nu_t \Rightarrow \nu_*$) as $t \to \infty$, it suffices to prove that $\langle \mathbf{1}, \mu_t \rangle \to 1$ and $r_t \to 0$ as $t \to \infty$.

Our main goal in this section is to establish these limits.

**Proposition 4.8.** *Suppose Assumptions 2.1 and 3.1(1) hold, and $\lambda > 1$. Then there exists $T \in (0,\infty)$ such that $B(t) = 1$ for all $t \geq T$. In addition,*

$$\langle \mathbf{1}, \mu_t \rangle \to 1 \quad \textit{and} \quad r_t \to 0, \quad \textit{as } t \to \infty, \tag{4.18}$$

*and also $\langle h, \nu_t \rangle \to 1$ as $t \to \infty$.*

To establish this proposition, we proceed in several steps, establishing various intermediate results in Steps 1–3, culminating in the proof of Proposition 4.8 in Step 4.

**Step 1.** We start with simple bounds on the measure $\theta_t$ defined in (4.13). Recall the convention $0 \log 0 = 0$.

**Lemma 4.9.** *We have* $\sup_t \langle h^s, \theta_t \rangle \le c_h$, $\int_0^\infty \langle h^s, \theta_t \rangle dt < \infty$, *and* $\int_0^\infty |\langle h^s, \theta_t \rangle \log \langle h^s, \theta_t \rangle| dt < \infty$.

*Proof.* Recall that $\theta_t = \nu_t - \mu_t$ is a nonnegative measure. Moreover, substituting $\psi = h^s$ in (4.13), we have for each $t > 0$,

$$\langle h^s, \theta_t \rangle = \int_{[0,\infty)} \frac{\bar{G}^s(x+t) h^s(x+t)}{\bar{G}^s(x)} \nu_0(dx). \tag{4.19}$$

For the first assertion, note that for all $t \ge 0$, $\langle h^s, \theta_t \rangle \le c_h \langle \mathbf{1}, \nu_t \rangle \le c_h$. The remaining claims will follow once we prove the following refinement of this bound, namely, for all $t \ge 0$,

$$\langle h^s, \theta_t \rangle \le c_h e^{-\varepsilon_h t}. \tag{4.20}$$

To see why this bound holds, first use the easily verifiable relation $\bar{G}^s(y) = e^{-\int_0^y h^s(u) du}$ and the definition of $\varepsilon_h$ to conclude that for all $x \ge 0$ and $t \ge 0$, $\bar{G}^s(x+t) \le \bar{G}^s(x) e^{-\varepsilon_h t}$. When substituted into (4.19), this yields

$$\langle h^s, \theta_t \rangle \le e^{-\varepsilon_h t} \int_{[0,\infty)} h^s(x+t) \nu_0(dx) \le c_h e^{-\varepsilon_h t} \langle \mathbf{1}, \nu_0 \rangle \le c_h e^{-\varepsilon_h t}.$$

This proves (4.20) and completes the proof. □

**Step 2.** We now obtain our main estimate on $r_t$ in Corollary 4.11, building off preliminary estimates obtained in Lemma 4.10. In what follows, we will say $(t_1, t_2) \subset [0, \infty)$ is a *busy interval* if $B_t = 1$ for $t \in (t_1, t_2)$, and say it is an *excursion interval* if $B_t < 1$ for $t \in (t_1, t_2)$ and $B_{t_1} = B_{t_2} = 1$.

Let $m(\cdot)$ denote the modulus of continuity of the continuous function $x \mapsto x \log x$ on the compact interval $[0, c_h]$. On $[0, e^{-1}]$ this function is decreasing. Now, for $0 \le x < y \le e^{-1}$, applying the inequality $p \log p + (1-p) \log(1-p) \le 0$ with $p = x/y$, we see that

$$0 \ge \frac{x}{y} \log \frac{x}{y} + \left(1 - \frac{x}{y}\right) \log \left(1 - \frac{x}{y}\right) = \frac{1}{y} [x \log x + (y-x) \log(y-x) - y \log y].$$

Hence, it follows that for $0 \le x < y \le e^{-1}$,

$$|x \log x - y \log y| = x \log x - y \log y \le (x-y) \log(y-x) = |(x-y) \log(y-x)|.$$

Moreover, in case $c_h > e^{-1}$, the function $x \mapsto x \log x$ is Lipschitz on $[e^{-1}, c_h]$. As a result, there is a constant $c_1$ (depending only on $c_h$) such that

$$m(x) \le |x \log x| + c_1 x, \qquad x \in [0, c_h]. \tag{4.21}$$

**Lemma 4.10.** *For $t \ge 0$, define $\Upsilon_t := m(\langle h^s, \theta_t \rangle)$, where $\theta_t$ is defined by (4.13), and $m$ is the modulus of continuity of $x \mapsto x \log x$, as defined above. If $(t_1, t_2)$ is a busy interval, then*

$$r_t \le r_{t_1} e^{-\varepsilon_h (t - t_1)} + \int_{t_1}^t \Upsilon_s ds, \quad t \in (t_1, t_2). \tag{4.22}$$

*On the other hand, if $(t_1, t_2)$ is an excursion, then*

$$r_{t_2} \le r_{t_1} + \int_{t_1}^{t_2} \Upsilon_s ds, \tag{4.23}$$

*and*

$$B'(t) = \lambda - \langle h^s, \nu_t \rangle, \qquad t \in (t_1, t_2). \tag{4.24}$$

*Furthermore, there exist finite positive constants $c_r$ and $c_{\text{lip}}$ such that $\sup_t |r_t| \le c_r$ and for any $0 \le s < t < \infty$, $|r_t - r_s| \le c_{\text{lip}} |t - s|$, showing that the function $t \to r_t$ is globally Lipschitz on $[0, \infty)$.*

*Proof.* Note that although the function $\tilde{f}$ defined in (4.14) is discontinuous in $t$ and in $x$, since $\bar{G}^s$ has a density, the relation (4.17) shows that $r_t$ is differentiable (although not continuously differentiable) with derivative

$$\frac{dr_t}{dt} = k_t \log k_t - \int_0^t g^s(t-x)k_x \log k_x dx = k_t \log k_t - \int_0^t g^s(x)k_{t-x} \log k_{t-x} dx. \tag{4.25}$$

Substituting the identities $g^s = h^s \bar{G}^s = h^s f^*$ and $k_{t-x} = \tilde{f}(x,t)/f^*(x)$ into (4.25), recalling the definition of $\tilde{f}$ from (4.14) recalling the convention that $0 \log 0 = 0$, and then applying Lemma 4.5, with $f$ replaced with $\tilde{f}(\cdot, t)$, we obtain

$$\begin{aligned} \frac{dr_t}{dt} &= k_t \log k_t - \int_0^\infty h^s(x)\tilde{f}(x,t) \log \frac{\tilde{f}(x,t)}{f^*(x)} dx \\ &\le k_t \log k_t - z_{\tilde{f}(\cdot,t)} \log z_{\tilde{f}(\cdot,t)} - \varepsilon_h r_t^+, \end{aligned} \tag{4.26}$$

where, as in Lemma 4.5, $z_{\tilde{f}(\cdot,t)} = \int_0^\infty h^s(x)\tilde{f}(x,t)dx$, which is equal to $\langle h^s, \mu_t \rangle$ by (4.13).

Now, suppose that $(t_1, t_2)$ is a busy interval for some $0 \le t_1 < t_2 \le \infty$. Then by (2.20) and Assumption 3.1(1), for $t \in (t_1, t_2)$, $c_h \ge k_t = \langle h^s, \nu_t \rangle = \langle h^s, \mu_t \rangle + \langle h^s, \theta_t \rangle$, which implies $k_t - z_{\tilde{f}(\cdot,t)} = \langle h^s, \theta_t \rangle \ge 0$. Since $m$ is the modulus of continuity of $x \mapsto x \log x$ on the interval $[0, c_h]$, it follows that

$$|k_t \log k_t - z_{\tilde{f}(\cdot,t)} \log z_{\tilde{f}(\cdot,t)}| \le m(\langle h^s, \theta_t \rangle) = \Upsilon_t.$$

When combined with (4.26), this shows that for any busy interval $(t_1, t_2)$,

$$\frac{dr_t}{dt} \le \Upsilon_t - \varepsilon_h r_t^+ \le \Upsilon_t - \varepsilon_h r_t, \quad t \in (t_1, t_2). \tag{4.27}$$

Now, let $\tilde{r}$ denote the solution to the differential equation $d\tilde{r}_t/dt = \Upsilon_t - \varepsilon_h \tilde{r}_t$ with the same initial condition as $r$, namely $\tilde{r}_{t_1} = r_{t_1}$. Then $\tilde{r}$ can be solved explicitly:

$$\tilde{r}_t = \tilde{r}_{t_1}^{-\varepsilon_h(t-t_1)} + \int_{t_1}^t e^{-\varepsilon_h(s-t_1)} \Upsilon_{t-s} ds \le r_{t_1} e^{-\varepsilon_h(t-t_1)} + \int_{t_1}^t \Upsilon_s ds, \quad t \in (t_1, t_2).$$

A simple comparison theorem for ordinary differential equations then shows that $r_t \le \tilde{r}_t$ for $t \in (t_1, t_2)$. This proves (4.22).

Next, consider an excursion interval $(t_1, t_2)$. Then (2.20) implies that $k_t = \lambda$ for $t \in (t_1, t_2)$. Moreover, it is not hard to see that the fluid age equation (2.7) holds with the test function $\varphi \equiv \mathbf{1}$, by approximating this function by compactly supported test functions whose derivatives in $x$ are bounded. Since $\varphi_x = \varphi_t = 0$, differentiating the equation yields (4.24). Toward showing (4.23), recall that $\langle \mathbf{1}, \nu_t \rangle = B(t)$ and $B(t_1) = B(t_2) = 1$ by definition of an excursion interval. Hence, it follows that

$$\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \langle h^s, \nu_t \rangle dt = \lambda. \tag{4.28}$$

Also, recalling $\theta_t = \nu_t - \mu_t$, we have

$$|\langle h^s, \mu_t \rangle \log \langle h^s, \mu_t \rangle - \langle h^s, \nu_t \rangle \log \langle h^s, \nu_t \rangle| \le m(\langle h^s, \nu_t \rangle - \langle h^s, \mu_t \rangle) = \Upsilon_t.$$

As a result, for $t \in (t_1, t_2)$, the right-hand side of (4.26) is bounded above by $\lambda \log \lambda - \langle h^s, \nu_t \rangle \log \langle h^s, \nu_t \rangle + \Upsilon_t$. Integrating both sides of (4.26) and using (4.28) and the convexity of $x \log x$, we have

$$r_{t_2} - r_{t_1} \le (t_2 - t_1)\lambda \log \lambda - \int_{t_1}^{t_2} \left( \langle h^s, \nu_t \rangle \log \langle h^s, \nu_t \rangle - \Upsilon_t \right) dt \le \int_{t_1}^{t_2} \Upsilon_t dt.$$

We now turn to the last assertion of the lemma. The bound $0 \le k_t \le c_h \vee \lambda$ implies that $|k_t \log k_t| \le c_2$ for some finite constant $c_2$. The boundedness of $t \to r_t$ thus follows from (4.17) and the fact that $\int_0^\infty \bar{G}^s(x)dx = 1$ (see Assumption 2.1(1)). By (4.25), the bound on $|k_t \log k_t|$ also implies that $dr_t/dt$ is bounded, and hence, that $t \mapsto r_t$ is globally Lipschitz on $[0, \infty)$. □

As a corollary, we obtain our main estimate on $r_t$. For $t > 0$, define

$$L(t) := \int_0^t 1_{\{B(s)=1\}} ds, \quad t > 0, \quad \text{and} \quad \mathcal{B} := \{t > 0 : B(t) = 1\}. \tag{4.29}$$

**Corollary 4.11.** *For every $s \geq 0$ and $t > s$, $t \in \mathcal{B}$,*

$$r_t \leq c_r e^{-\varepsilon_h (L(t)-L(s))} + \int_s^t \Upsilon_\tau d\tau, \tag{4.30}$$

*where $c_r$ is the constant from Lemma 4.10.*

*Proof.* Fix $s \geq 0$ and $t > s$, $t \in \mathcal{B}$. Denote $t_0 := \inf\{u \geq s : B_u = 1\}$. Fix a nonempty open interval $(s_0, s_1) \subset (t_0, t)$. Then $(s_0, s_1)$ is said to be a *maximal* busy interval if it is a busy interval that is not a proper subset of any open busy interval contained in $(t_0, t)$. Further, $(s_0, s_1) \subset (t_0, t)$ is referred to as *admissible* if it is either an excursion or a maximal busy interval. Since $B$ is continuous it is clear that $\mathcal{O} := \{u \in (s, t) : B_u < 1\}$ is an open set, and hence, can be written as a countable union of open intervals. Thus, there are at most a countable number of excursions. Since any maximal busy interval must be contiguous to one of the intervals comprising $\mathcal{O}$, it follows that the collection of admissible intervals is also countable. For $u > 0$, define a *$u$-admissible* interval to be an admissible interval whose length is at least $u$. Denote by $\mathcal{T}_u$ the complement in $(t_0, t)$ of the union of all $u$-admissible intervals. Then, as $u \to 0$, the Lebesgue measure $|\mathcal{T}_u|$ of this set clearly converges to zero.

Let $u > 0$ be given, and let $I_u$ be the number of $u$-admissible intervals. Since there are only a finite number of such intervals, we can label the intervals $(t_n, t'_n)$, $n = 1, \ldots, I_u$ in such a way that $s \leq t_0 \leq t_1 < t_2 < t_3 < \cdots < t_{I_u} \leq t$. Let $c_{\text{lip}}$ denote the (global) Lipschitz constant of $t \mapsto r_t$, which exists by Lemma 4.10. We now show by induction that, for $n = 1, 2, \ldots, I_u$,

$$r_{t_n} \leq r_{t_1} e^{-\varepsilon_h \sum_{i=1}^{n-1} (L(t'_i)-L(t_i))} + \sum_{i=1}^{n-1} \int_{t_i}^{t'_i} \eta_\tau d\tau + c_{\text{lip}} \sum_{i=1}^{n-1} (t_{i+1} - t'_i), \tag{4.31}$$

where a sum with the upper limit less than the lower limit is taken to be zero.

*Base Case:* For $n = 1$, (4.31) reduces to the trivial inequality $r_{t_1} \leq r_{t_1}$, and thus is satisfied.

*Induction step:* Assuming (4.31) holds for an arbitrary $n \in \{1, \ldots, I_u - 1\}$, we show it holds for $n+1$. From (4.22) and (4.23) of Lemma 4.10, along with the fact that $L(t'_n) - L(t_n)$ is equal to zero if $(t_n, t'_n)$ is an excursion, and is equal to $t'_n - t_n$ if it is a busy interval, we have

$$r_{t'_n} \leq r_{t_n} e^{-\varepsilon_h (L(t'_n)-L(t_n))} + \int_{t_n}^{t'_n} \Upsilon_\tau d\tau.$$

Using this estimate, the Lipschitz continuity of $r_t$ established in Lemma 4.10 and the induction hypothesis, it follows that

$$\begin{aligned}
r_{t_{n+1}} &\leq r_{t'_n} + c_{\text{lip}}(t_{n+1} - t'_n) \\
&\leq r_{t_n} e^{-\varepsilon_h (L(t'_n)-L(t_n))} + \int_{t_n}^{t'_n} \Upsilon_\tau d\tau + c_{\text{lip}}(t_{n+1} - t'_n) \\
&\leq \Big( r_{t_1} e^{-\varepsilon_h \sum_{i=1}^{n-1} (L(t'_i)-L(t_i))} + \sum_{i=1}^{n-1} \int_{t_i}^{t'_i} \Upsilon_\tau d\tau + c_{\text{lip}} \sum_{i=1}^{n-1} (t_{i+1} - t'_i) \Big) e^{-\varepsilon_h (L(t'_n)-L(t_n))} \\
&\qquad + \int_{t_n}^{t'_n} \Upsilon_\tau d\tau + c_{\text{lip}}(t_{n+1} - t'_n) \\
&\leq r_{t_1} e^{-\varepsilon_h \sum_{i=1}^{n} (L(t'_i)-L(t_i))} + \sum_{i=1}^{n} \int_{t_i}^{t'_i} \Upsilon_\tau d\tau + c_{\text{lip}} \sum_{i=1}^{n} (t_{i+1} - t'_i).
\end{aligned}$$

This proves (4.31) by induction.

Next, note that each of the intervals $(t_0, t_1)$, $(t'_i, t_{i+1})$ for $i = 1, \ldots, I_u - 1$, and $(t_{I_u}, t)$, is a subset of $\mathcal{T}_u$. Hence, we have

$$L(t) - L(t_0) - \sum_{i=1}^{I_u-1} (L(t'_i) - L(t_i)) \leq (L(t_1) - L(t_0)) + \sum_{i=1}^{I_u-1} (L(t_{i+1}) - T(t'_i)) + (L(t) - L(t_{I_u})) \leq |\mathcal{T}_u|.$$

Hence, on applying (4.31) with $n = I_u$, noting that the last term on the right-hand side is bounded by $c_{\text{lip}}|\mathcal{T}_u|$, we obtain on taking $u \to 0$, $r_t \leq c_r e^{-\varepsilon_h(L(t)-L(t_0))} + \int_{t_0}^t \Upsilon_\tau d\tau$. Finally, since by definition $s \leq t_0$ and $L(t_0) = L(s)$, the lemma follows. □

**Step 3.** We now prove that $\bar{L} := \sup_t L(t) = \infty$. Note that this implies that the "servers" become busy infinitely often, as one might expect in the supercritical regime $\lambda > 1$. (We will later use this to prove the stonger condition that the complement of $\mathcal{B}$ is bounded.)

Arguing by contradiction, assume that $\bar{L} < \infty$. By (4.13), clearly $\langle \mathbf{1}, \mu_t \rangle \leq \langle \mathbf{1}, \nu_t \rangle \leq 1$ and, by Assumption 3.1(1), $\langle h, \nu_t \rangle \leq c_h$, for all $t > 0$. However, (4.13), (4.14) and (2.20) together implly

$$\begin{aligned}
\langle \mathbf{1}, \mu_t \rangle = \int_0^\infty \tilde{f}(x,t)dx &= \int_0^t \bar{G}^s(x) k_{t-x} dx \\
&= \int_0^t \bar{G}^s(x)[\lambda 1_{\{B_{t-x}<1\}} + \langle h^s, \nu_{t-x} \rangle 1_{\{B_{t-x}=1\}}]dx \\
&\geq \lambda \int_0^t \bar{G}^s(x)dx - \lambda \int_0^t \bar{G}^s(x) 1_{\{B_{t-x}=1\}} dx.
\end{aligned}$$

Moreover, it is also true that

$$\begin{aligned}
\int_0^t \bar{G}^s(x) 1_{\{B_{t-x}=1\}} dx &\leq \int_0^{t/2} 1_{\{B_{t-x}=1\}} dx + \int_{t/2}^t \bar{G}^s(x)dx \\
&\leq (L(t) - L(t/2)) + \int_{t/2}^\infty \bar{G}^s(x)dx.
\end{aligned}$$

Recalling that $\int_0^\infty \bar{G}^s(x)dx = 1$ (see Assumption 2.1), if $\bar{L} < \infty$ the above expression converges to zero as $t \to \infty$. Hence, $\liminf_{t\to\infty} \langle \mathbf{1}, \mu_t \rangle \geq \lambda > 1$, which is a contradiction. This proves $\bar{L} = \infty$.

**Step 4.** We now combine the above results to prove Proposition 4.8.

*Proof of Proposition 4.8.* We first claim that to establish (4.18), it suffices to show that $B(t) = 1$ for all sufficiently large $t$. Recalling that $\Upsilon = m(\langle h^s, \theta \rangle)$ is integrable on $[0, \infty)$ by Lemma 4.9 and the bound (4.21) on $m$, and that $0 \leq L(t) \to \infty$ as $t \to \infty$ by Step 3, which implies $\mathcal{B}$ is unbounded, we can send first $t \to \infty$ along $\mathcal{B}$ and then $s \to \infty$ in (4.30) of Corollary 4.11, to obtain $\limsup_{t\to\infty,\ t\in\mathcal{B}} r_t \leq 0$. We cannot directly deduce from this that the limit of $r_t$ along $\mathcal{B}$ is zero, since $r_t = R(\mu_t \| \nu_*)$ could be negative. However, for $t \in \mathcal{B}$, $B(t) = \langle \mathbf{1}, \nu_t \rangle = 1$ and hence, $\langle \mathbf{1}, \mu_t \rangle = 1 - \langle \mathbf{1}, \theta_t \rangle$. Since $r_t = R(\mu_t \| \nu_*)$, and $\langle \mathbf{1}, \theta_t \rangle \to 0$ by Remark 4.7, when combined with (4.6) this implies $\limsup_{t\to\infty, t\in\mathcal{B}} r_t \geq \limsup_{t\to\infty, t\in\mathcal{B}} \langle \mathbf{1}, \mu_t \rangle \ln \langle \mathbf{1}, \mu_t \rangle = 0$. Hence,

$$\lim_{t\to\infty,\ t\in\mathcal{B}} \langle \mathbf{1}, \mu_t \rangle = 1 \quad \text{and} \quad \lim_{t\to\infty,\ t\in\mathcal{B}} r_t = 0, \tag{4.32}$$

If $\mathcal{B} \supseteq [t_0, \infty)$ for some finite $t_0$, this clearly proves (4.18), and the claim follows.

We now turn to the proof of the fact that $B(t) = 1$ outside a finite interval. First note that, (4.32) and the Pinsker-type inequality (4.15) together show that

$$\lim_{t\to\infty,\ t\in\mathcal{B}} \int_0^\infty |\tilde{f}(x,t) - f^*(x)|dx = 0.$$

Thus, given $\varepsilon_0 := \frac{\lambda-1}{4}$ there exists $T \in \mathcal{B}$ such that

$$c_h\langle \mathbf{1}, \theta_t\rangle < \varepsilon_0 \quad \text{and} \quad c_h \int_0^\infty |\tilde{f}(x,t) - f^*(x)|dx < \varepsilon_0 \qquad \text{for all } t \geq T,\ t \in \mathcal{B}. \tag{4.33}$$

We claim that $[T,\infty) \subset \mathcal{B}$. Arguing by contradiction, assume there exists $T' > T$ for which $T' \notin \mathcal{B}$, that is, such that $B(T') < 1$. Let $\tau := \sup\{t < T' : B_t = 1\}$. By the continuity of $B$, $T \leq \tau < T'$ and $\tau \in \mathcal{B}$; in particullar, the estimates in (4.33) are valid for $t = \tau$. Find $t^* > 0$ so small that $G^s(t^*) < \frac{1}{4}$ and $0 < t^* < T' - \tau$. For all $0 \leq t \leq t^*$, applying Lemma 2.7 with $T = \tau$ and (2.19) with $\psi = h^s$, and using the fact that (2.20) implies $K(\tau + t) - K(\tau) = \lambda t$ for $t \in (0, t^*)$, the identity $g^s = h^s\bar{G}^s$, the upper bound on $h^s$ from Assumption 3.1(1) and (4.33), we obtain

$$\begin{aligned}
\langle h^s, \nu_{\tau+t}\rangle &= \int_0^\infty \frac{g^s(x+t)}{\bar{G}^s(x)}\nu_\tau(dx) + \lambda\int_0^t g^s(t-s)ds \\
&= \int_0^\infty \frac{g^s(x+t)}{\bar{G}^s(x)}\theta_\tau(dx) + \int_0^\infty \frac{g^s(x+t)}{\bar{G}^s(x)}\nu^*(dx) + \int_0^\infty \frac{g^s(x+t)}{\bar{G}(x)}(\mu_\tau(dx) - \nu^*(dx)) + \lambda\int_0^t g^s(t-s)ds \\
&\leq c_h\langle \mathbf{1}, \theta_\tau\rangle + \int_0^\infty g^s(x+t)dx + c_h\int_0^\infty |f(\tau,x) - f^*(x)|dx + \lambda G^s(t) \\
&\leq \varepsilon_0 + 1 - G^s(t) + \varepsilon_0 + \lambda G^s(t) \\
&= 1 + (\lambda-1)G^s(t) + 2\varepsilon_0 \leq 1 + 3\varepsilon_0 = \lambda - \varepsilon_0.
\end{aligned}$$

Thus for all $\tau < s < \tau + t^*$, $\langle h, \nu_s\rangle \leq \lambda - \varepsilon_0$. Next, since the interval $(\tau, \tau + t^*)$ is a subset of an excursion, equation (4.24) for $B$ is valid for $t$ in that interval, and it follows that $B'(s) \geq \varepsilon_0$ for $s \in (\tau, \tau + t^*)$. Bsy the continuity of $B$,

$$B(s) \geq 1 + (t-\tau)\varepsilon_0 > 1, \qquad s \in (\tau, \tau + t^*),$$

which is a contradiction. We have thus shown that $B(t) = 1$ for all sufficiently large $t$. Together with (4.32), this proves (4.18).

To conclude the proof of the proposition, it only remains to show that $\langle h^s, \nu_t\rangle \to 1$ as $t \to \infty$. Fix $T \in (0,\infty)$ such that $B(t) = 1$ for all $t \geq T$. Then using Lemma 2.7 and equation (2.19) with $\psi(x) = h^s(x)$, and noting from (2.20) that $K'(T+s) = \langle h, \nu_{T+s}\rangle$ for all $s > 0$, and recalling again that $g^s = \bar{G}^s h^s$, we have

$$\langle h^s, \nu_{T+s}\rangle = z(s) + \int_{[0,s]} g^s(T+s-w)\langle h^s, \nu_{T+w}\rangle dw,$$

where $z(s) := \int_{[0,\infty)} \frac{g^s(x+s)}{\bar{G}^s(x)}\nu_T(dx)$. Next, note that $g^s(s) = h^s(s)\bar{G}^s(s) \leq c_h\bar{G}^s(s)$. Since $\bar{G}^s$ is decreasing and integrable over $[0,\infty)$, it is also directly Riemann integrable (see Prop. 2.16(c), Ch. 9 of [14]), and thus, so is $g^s$. Hence, by the key renewal theorem (Theor. 2.8, Ch. 9 of [14]), $\langle h^s, \nu_{T+s}\rangle$ converges as $s \to \infty$ to $\int_0^\infty z(s)ds / \int_0^\infty xg^s(x)dx = \int_0^\infty z(s)ds$, since by Assumption 2.1, $\int_0^\infty xg^s(x)dx = 1$. Thus,

$$\begin{aligned}
\lim_{s\to\infty}\langle h^s, \nu_{T+s}\rangle &= \int_0^\infty \int_{[0,\infty)} \frac{g^s(x+s)}{\bar{G}^s(x)}\nu_T(dx)ds \\
&= \int_{[0,\infty)} \frac{1}{\bar{G}^s(x)}\Big(\int_0^\infty g^s(x+s)ds\Big)\nu_T(dx) = \int_{[0,\infty)} \nu_T(dx),
\end{aligned}$$

which is equal to 1 by our choice of $T$. This completes the proof of the proposition. □

4.3. **Proof of Convergence when the Hazard Rate Function is Decreasing.** In this section, we assume throughout that Assumption 2.1 and Assumption 3.1(2) hold, and we establish Theorem 3.2(2) in this case, as well as Theorem 3.2(3). In addition, fix $\lambda \geq 1$, and suppose that $(X, \nu, \eta)$ is the solution to the fluid equations with arrival rate $\lambda$ and some initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$. Also, recall from (2.17) that $B(t) = \langle \mathbf{1}, \nu_t\rangle$, and define

$$W(t) := B(t) - \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)}\nu_0(dx), \quad t \geq 0. \tag{4.34}$$

Note that $W(t)$ represents the fluid mass of jobs that arrived after time 0 and are still in service at time $t$.

We will first establish the following key result.

**Proposition 4.12.** *Suppose Assumption 2.1 and Assumption 3.1(2) hold, and $\lambda \geq 1$. Then we have*

$$\lim_{t\to\infty} W(t) = \lim_{t\to\infty} B(t) = 1, \tag{4.35}$$

*Further, if $\lambda > 1$, there exists $T \in [0,\infty)$ such that $B(t) = 1$ for all $t \geq T$.*

Before launching into the proof, we derive some useful relations. Setting $\psi \equiv \mathbf{1}$ in (2.19) and using integration by parts, it follows that

$$\begin{aligned} B(t) = \langle \mathbf{1}, \nu_t \rangle &= \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + \int_0^t \bar{G}^s(t-s) dK(s) \\ &= \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + K(t) - \int_0^t K(s) g^s(t-s) ds, \end{aligned} \tag{4.36}$$

which when rearranged yields

$$K(t) = W(t) + \int_0^t K(t-s) g^s(s) ds. \tag{4.37}$$

Then, (4.36), (4.34) and the fact that $\nu_t$ is a sub-probability measure, together imply that for each $t \geq 0$,

$$W(t) = \int_0^t \bar{G}^s(t-s) dK(s) \geq 0 \quad \text{and} \quad W(t) \leq B(t) \leq 1. \tag{4.38}$$

Together with (4.37) and the renewal theorem (see Chapter V of [6]), this implies

$$K(t) = W(t) + Z(t), \qquad \text{with} \quad Z(t) := \int_0^t W(t-s)\, dU_s(s), \tag{4.39}$$

and $U_s$ is equal to the renewal function of the distribution with density $g^s$. Now, (2.8) implies that

$$D(t) := \int_0^t \langle h, \nu_s \rangle ds = B(0) - B(t) + K(t), \qquad t \geq 0.$$

Then by (4.39), (4.34) and (4.36), we obtain

$$\begin{aligned} D(t) &= \langle \mathbf{1}, \nu_0 \rangle - \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + Z(t) \\ &= \int_{[0,H^s)} \frac{G^s(x+t) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) + Z(t). \end{aligned} \tag{4.40}$$

Under Assumption 3.1(2), the hazard rate function $h^s$ is decreasing and hence, by Theorem 3 of [12], the renewal function $U_s$ is concave. Since $G^s$ has density $g^s$, the density $u_s := U_s'$ exists by Proposition 2.7 of [6] and $u_s(x) = \sum_{n=1}^{\infty} (g^s)^{\star n}(x)$, $x \geq 0$, which in particular implies that $u_s(0) = g^s(0)$. Moreover, by Alexandrov's Theorem (cf. page 172 of [34]), the concavity of $U_s$ implies that $u_s$ is non-increasing, that is,

$$u_s'(t) \leq 0, \qquad \text{for a.e. } t \geq 0. \tag{4.41}$$

Now, differentiation of both sides of the defining equation for $Z$ in (4.39) yields

$$Z'(t) = W(t) u_s(0) + \int_0^t W(t-s) u_s'(s) ds, \qquad \text{for a.e. } t \geq 0. \tag{4.42}$$

On the other hand, differentiating the equation for $K$ in (4.39) and using (2.20), one obtains, for a.e. $t \geq 0$,

$$\begin{aligned} W'(t) &= K'(t) - Z'(t) \\ &= \begin{cases} \lambda - Z'(t) & \text{if } B(t) < 1, \\ D'(t) - Z'(t) & \text{if } B(t) = 1 \text{ and } Q(t) > 0, \\ \lambda \wedge D'(t) - Z'(t) & \text{if } B(t) = 1 \text{ and } Q(t) = 0. \end{cases} \end{aligned} \tag{4.43}$$

Next, differentiating both sides of (4.40), we obtain for a.e. $t \geq 0$,

$$D'(t) = \int_{[0,H^s)} \frac{g^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + Z'(t) \geq Z'(t).$$

Therefore, by (4.43), for a.e. $t \geq 0$,

$$Z'(t) \leq \lambda \quad \Rightarrow \quad W'(t) \geq 0. \tag{4.44}$$

We now establish some auxiliary results that will be used in the proof of Proposition 4.12.

**Lemma 4.13.** *Suppose $\lambda \geq 1$. Then there is no $T \in (0,\infty)$ and $c \in (0,1)$ such that $W(t) < c$ for all $t \geq T$. The same assertion also holds when $W$ is replaced with $B$.*

*Proof.* Suppose the statement of the lemma is not true, that is, suppose there exists $T > 0$ and $c \in (0,1)$ such that $W(t) < c$ for all $t \geq T$. Since $\int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) \to 0$ as $t \to \infty$, by (4.34), there exists $T' > T$ such that $B(t) < 1$ for all $t \geq T'$. In turn, by (2.8), it follows that $K'(t) = \lambda$, for all $t \geq T'$, and hence (4.36) and (4.34) imply that

$$W(t) = \int_0^t \bar{G}^s(t-s)dK(s) = \int_0^{T'} \bar{G}^s(t-s)dK(s) + \lambda \int_{T'}^t \bar{G}^s(t-s)ds.$$

As $t \to \infty$, the first term converges to zero by the dominated convergence theorem and the pointwise limit $\bar{G}^s(t-s) \to 0$. For the same reason, the second term converges to $\lim_{t\to\infty} \lambda \int_0^t \bar{G}^s(t-s)ds = \lambda \int_0^\infty \bar{G}^s(s)ds$, which is equal to $\lambda$ by (2.1) of Assumption 2.1. Thus, $\lim_{t\to\infty} W(t) \geq \lambda$, which is a contradiction, thus proving the first assertion of the lemma. Since, by (4.34), $B(t) - W(t) = \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) \to 0$ as $t \to \infty$, the same assertion holds also for $B$. □

Next, substituting into (4.42) the inequality (4.41), the relation $u_s(0) = g^s(0)$ and the fact that $W(t) \in [0,1]$ for each $t \geq 0$ due to (4.38), we see that

$$Z'(t) \leq W(t)u_s(0) = W(t)g^s(0) \leq g^s(0) \quad \text{for a.e. } t \geq 0. \tag{4.45}$$

We also observe that since the hazard rate function $h^s$ is decreasing by Assumption 3.1(2), then $g^s(0) > 0$. (Otherwise, if $g^s(0) = 0$, then $h^s(0) = 0$, which implies that $0 \leq h^s(t) \leq h^s(0) = 0$ for each $t \geq 0$ and thus, $g^s(t) = 0$ for all $t \geq 0$, which would contradict the fact that $g^s$ is the density of $G^s$.) Therefore, for $n \in \mathbb{N} \cup \{0\}$ and $\varepsilon \in (0, \frac{1}{2})$, define

$$\lambda_n := \frac{\lambda - \varepsilon}{g^s(0)} \left( \sum_{i=0}^n \left(1 - \frac{1}{g^s(0)}\right)^i \right) = \frac{\lambda - \varepsilon}{g^s(0)} \frac{1 - \left(1 - \frac{1}{g^s(0)}\right)^{n+1}}{1 - \left(1 - \frac{1}{g^s(0)}\right)} = (\lambda - \varepsilon)\left(1 - \left(1 - \frac{1}{g^s(0)}\right)^{n+1}\right), \tag{4.46}$$

and

$$\tau_n := \sup\{t > 0 : W(t) < \lambda_n\}. \tag{4.47}$$

If $\tau_n < \infty$, then

$$W(\tau_n + t) \geq \lambda_n \quad \forall t \geq 0. \tag{4.48}$$

**Lemma 4.14.** *Suppose $\lambda \geq 1$, $\varepsilon \in (0, \frac{1}{2})$ and $g^s(0) > \lambda - \varepsilon$. Then $\tau_n < \infty$ and hence,* (4.48) *holds for all $n \in \mathbb{N}$ with $n < n^*$, where $n^* := \sup\{n \in \mathbb{N}_0 : \lambda_n < 1\}$, and also for $n = n^*$, if $n^* < \infty$.*

*Proof.* Since $g^s(0) > \lambda - \varepsilon > \frac{1}{2}$ by the assumptions of the lemma, it follows that $|1 - \frac{1}{g^s(0)}| < 1$ and (4.46) then implies that

$$\lambda_n \to \lambda - \varepsilon \text{ as } n \to \infty. \tag{4.49}$$

We prove the lemma by induction. We first start with the base case $n = 0$, where $\lambda_0 = (\lambda - \varepsilon)/g^s(0)$. Note that $\lambda_0 < 1$ by the assumptions of the lemma. We argue by contradiction to show that

$$W(t) < \frac{\lambda - \varepsilon}{g^s(0)} \quad \text{for all } t \in (0, \tau_0). \tag{4.50}$$

Note that (4.50) holds trivially if $\tau_0 = 0$. So, suppose $\tau_0 > 0$ and (4.50) does not hold. Then there must exist $0 < t_1 < \tau_0$ for which $W(t_1) \geq \frac{\lambda-\varepsilon}{g^s(0)}$. It follow from (4.45) and (4.44), that for a.e. $t \in (t_1, \tau_0)$, the inequality $W(t) < \frac{\lambda-\varepsilon}{g^s(0)}$ implies that $Z'(t) \leq W(t)g^s(0) < \lambda - \varepsilon < \lambda$ and hence $W'(t) \geq 0$. Since $W$ is absolutely continuous, $W(t) \geq \frac{\lambda-\varepsilon}{g^s(0)} = \lambda_0$ for all $t \in [t_1, \tau_0)$ (see Lemma B.1). This contradicts the definition of $\tau_0$, and thus, (4.50) holds. If $\tau_0 = \infty$, then (4.50) implies $W(t) < \lambda_0 < 1$ for all $t > 0$, which contradicts Lemma 4.13. Thus, $\tau_0 < \infty$. This completes the proof of the base case.

Now, suppose that $\tau_k < \infty$ for some $k \in \mathbb{N} \cup \{0\}$, with $k < n^*$ if $n^* < \infty$. It follows that $\lambda_{k+1} < 1$ by the choice of $k$ and the definition of $n^*$. By the definition of $\tau_k$ and the continuity of $W$,

$$W(\tau_k + t) \geq \lambda_k \text{ for all } t \in [0, \infty). \tag{4.51}$$

Then for a.e. $t \geq 0$, by (4.42), (4.41), (4.51) and the relations $W(t) \geq 0$ and $u_s(0) = g^s(0)$, we have

$$Z'(\tau_k + t) = W(\tau_k + t)g^s(0) + \int_0^t W(\tau_k + t - s)u_s'(s)ds + \int_t^{\tau_k+t} W(\tau_k + t - s)u_s'(s)ds \tag{4.52}$$

$$< W(\tau_k + t)g^s(0) + \lambda_k \left(u_s(t) - g^s(0)\right). \tag{4.53}$$

Since Assumption 3.1(2) implies that the integrable function $g^s$ is also bounded, it lies in $\mathbb{L}^{1+\varepsilon}(0, \infty)$ for any $\varepsilon > 0$, and satisfies $g^s(t) \to 0$ as $t \to \infty$. Thus, by Theorem 12 of [39] we can conclude that $\lim_{t\to\infty} u_s(t) = 1$. Hence, there exists $\sigma_k > 0$ such that

$$(\lambda - \varepsilon) + \lambda_k(u_s(t) - 1) = (\lambda - \varepsilon) + \lambda_k \left(g^s(0) - 1\right) + \lambda_k \left(u_s(t) - g^s(0)\right) < \lambda \text{ for all } t \geq \sigma_k. \tag{4.54}$$

We now show that the following statement cannot hold:

$$W(\tau_k + t) < \lambda_{k+1} = \frac{\lambda - \varepsilon}{g^s(0)} + \lambda_k \left(1 - \frac{1}{g^s(0)}\right) \text{ for all } t > \sigma_k, \tag{4.55}$$

where the equality follows from (4.46). Indeed, if this were true, then this would imply that $W(t) < \lambda_{k+1} < 1$ for all $t \geq \sigma_k' := \tau_k + \sigma_k$, which contradicts Lemma 4.13. Thus, (4.55) does not hold or, in other words, there exists $\tau_k' \in (\sigma_k', \infty)$ such that $W(\tau_k') \geq \lambda_{k+1}$. We now show that for a.e. $t \in (0, \infty)$, if $W(\tau_k' + t) < \lambda_{k+1}$ then $W'(\tau_k' + t) \geq 0$. Indeed, if the first inequality is true, then substituting this into (4.53) with $\tau_k'$ in place of $\tau_k$, and using (4.54), it follows that $Z'(\tau_k' + t) < \lambda$. When combined with (4.44) the latter implies $W'(\tau_k' + t) \geq 0$. Hence (applying Lemma B.1 with $f = W$, $c = \lambda_{k+1}$, $T = \tau_k'$, $S = \infty$), it follows that $W(\tau_k' + t) \geq \lambda_{k+1}$ for all $t \geq 0$, thus showing that $\tau_{k+1} \leq \tau_k' < \infty$. By induction, it follows that for each $0 \leq n < n^*$, $\tau_n < \infty$ and hence, (4.48) holds, and if $n^* < \infty$ then also $\tau_{n^*} < \infty$ and (4.48) holds with $n = n^*$. This completes the proof of the lemma.

□

We are now in a position to present the proof of Proposition 4.12.

*Proof of Proposition 4.12.* We first prove the proposition when $\lambda = 1$. For this, we consider two cases.
*Case 1a:* $g^s(0) \leq 1$. In this case, (4.45) shows that $Z'(t) \leq 1$ for a.e. $t \geq 0$, then (4.44) implies that for a.e. $t \geq 0$, $W'(t) \geq 0$. Since $W$ is absolutely continuous by (4.38) this implies that $W$ is increasing on $[0, \infty)$ and $b := \lim_{t\to\infty} W(t)$ exists. Furthermore, (4.34) and the fact that $\bar{G}^s(x + t) \to 0$ as $t \to \infty$ for every $x \in [0, H^s)$, imply $b = \lim_{t\to\infty} B(t)$. We now argue by contradiction to show that $b = 1$. Suppose $b < 1$, then for any $T < \infty$, there exists $T_1 < \infty$ such that for $t \geq 0$, $B(T_1 + t) < 1$ and thus, by (2.20) $K'(T_1 + t) = 1$. Now, recalling $B(\cdot) = \langle \mathbf{1}, \nu_\cdot \rangle$ from (4.36) and combining Lemma 2.7 and Theorem 2.6, it follows that (2.19)

holds with $\psi = \mathbf{1}$, and $\nu_t$ and $K_t$ replaced with $\nu_{T_1+t}$, and $K_{T_1+t} - K_{T_1}$, respectively, or in other words, for each $t \geq 0$,

$$B(T_1+t) \quad = \quad \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_{T_1}(dx) + \int_0^t \bar{G}^s(t-s) K'(T_1+s) ds.$$

When combined with the relation $K'(T_1+\cdot) = \lambda = 1$ a.e., this implies that for each $t \geq 0$.

$$B(T_1+t) \quad = \quad \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_{T_1}(dx) + \int_0^t \bar{G}^s(t-s) ds,$$

Sending $t \to \infty$, using $\bar{G}^s(x+t) \to 0$ pointwise and the dominated convergence theorem, as well as (2.2) of Assumption 2.1, this implies $b = \lim_{t\to\infty} B(t) = 1$. This contradicts the supposition that $b < 1$, and thus proves that $b = 1$.

*Case 1b:* $g^s(0) > 1$. In this case, by (4.46),

$$\lambda_n = (1-\varepsilon)\left(1 - \left(1 - \frac{1}{g^s(0)}\right)^{n+1}\right) < 1 \text{ for all } n \geq 1.$$

Thus, by Lemma 4.14, for each $n \geq 1$, we have $\tau_n < \infty$ and so (4.48) implies $\liminf_{t\to\infty} W(t) \geq \lambda_n$ for each $n \geq 1$. By (4.49), we obtain $\liminf_{t\to\infty} W(t) \geq 1-\varepsilon$. Sending $\varepsilon \downarrow 0$, we obtain $\liminf_{t\to\infty} W(t) \geq 1$. Since $\limsup_{t\to\infty} W(t) \leq 1$ by (4.38) it follows that in fact $\lim_{t\to\infty} W(t) = 1$. When combined with (4.34) and the fact that $\bar{G}^s(x+t) \to 0$ as $t \to \infty$ for every $x \in [0, H^s)$, it follows that $\lim_{t\to\infty} B(t) = 1$, thus proving the proposition in this case.

We next prove the proposition for the case that $\lambda > 1$. Let $\varepsilon > 0$ be small enough such that $\lambda - \varepsilon > 1$. We now consider two cases.

*Case 2a:* $g^s(0) \leq \lambda - \varepsilon$. In this case, (4.45) shows that $Z'(t) \leq \lambda - \varepsilon < \lambda$ for a.e. $t \geq 0$, and hence, (4.44) implies that for a.e. $t \geq 0$, $W'(t) \geq 0$. Moreover, by (4.43), we have $W'(t) = \lambda - Z'(t) \geq \varepsilon$ if $B(t) < 1$. By the definition of $W$ in (4.34), we obtain

$$B'(t) = W'(t) + \int_{[0,H^s)} \frac{g^s(x+t)}{\bar{G}^s(x)} \nu_0(dx).$$

Since $h^s$ is decreasing, we have $h^s(x+t) \leq h^s(0)$ for each $x \in [0, H^s - t)$, and an application of the dominated convergence theorem shows that

$$\int_{[0,H^s)} \frac{g^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) \leq \int_{[0,H^s)} \frac{h^s(0)\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) \to 0 \text{ as } t \to \infty.$$

The last three displays together imply that there exists $T \in (0,\infty)$ such that $B'(t) > \varepsilon/2$ whenever $B(t) < 1$ for a.e. $t \in [T, \infty)$. Since $B$ is bounded (by 1), the inequality $B(t) < 1$ cannot hold for all $t \geq T$. In other words, there must exist $T' > T$ such that $B(T') = 1$. Since $B$ is absolutely continuous and bounded by 1 (applying Lemma B.1 with $f = B$, $c = 1$, $T = T'$ and $S = \infty$), we conclude that $B(t) = 1$ for all $t \in [T', \infty)$.

*Case 2b:* $g^s(0) > \lambda - \varepsilon$. Then $n^* < \infty$ since (4.46) shows that $\lambda_n \uparrow (\lambda - \varepsilon) > 1$ as $n \to \infty$. Since by Lemma 4.14, $\tau_{n^*} < \infty$, then the continuity of $W$ dictates that $W(\tau_{n^*}) = \lambda_{n^*}$. Together with (4.53) with $k = n^*$ and the fact that $W$ is bounded by 1 due to (4.38), this implies that for a.e. $t \geq 0$,

$$\begin{aligned} Z'(\tau_{n^*}+t) \quad &\leq \quad W(\tau_{n^*}+t) g^s(0) + \lambda_{n^*}\left(u_s(t) - g^s(0)\right) \\ &\leq \quad (1-\lambda_{n^*}) g^s(0) + \lambda_{n^*} u_s(t). \end{aligned}$$

By the definition of $n^*$, we have $\lambda_{n^*} < 1 \leq \lambda_{n^*+1}$. Together with the definition of $\lambda_n$ in (4.46), this implies that

$$1 - \lambda_{n^*} \leq \lambda_{n^*+1} - \lambda_{n^*} = \frac{\lambda - \varepsilon}{g^s(0)} \left(1 - \frac{1}{g^s(0)}\right)^{n^*+1}.$$

Combining the above two displays, we obtain

$$Z'(\tau_{n^*}+t) \le (\lambda-\varepsilon)\left(1-\frac{1}{g^s(0)}\right)^{n^*+1} + \lambda_{n^*}u_s(t).$$

Recalling that $\lim_{t\to\infty} u_s(t) = 1$ and using the expression for $\lambda_{n^*}$ from (4.46), it follows that as $t\to\infty$,

$$(\lambda-\varepsilon)\left(1-\frac{1}{g^s(0)}\right)^{n^*+1} + \lambda_{n^*}u_s(t) \to (\lambda-\varepsilon)\left(1-\frac{1}{g^s(0)}\right)^{n^*+1} + \lambda_{n^*} = \lambda-\varepsilon.$$

Thus, for all $t$ large enough, $Z'(\tau_{n^*}+t) < \lambda-\varepsilon/2$. However, note that by (4.43), we have

$$W'(\tau_{n^*}+t) = \lambda - Z'(\tau_{n^*}+t) > \varepsilon/2 \text{ if } B(\tau_{n^*}+t) < 1.$$

By the definition of $W$ in (4.34), we obtain

$$B'(\tau_{n^*}+t) = W'(\tau_{n^*}+t) + \int_{[0,H^s)} \frac{g^s(x+\tau_{n^*}+t)}{\bar{G}^s(x)}\nu_0(dx).$$

Since $h^s$ is decreasing, it follows that $h^s(x+\tau_{n^*}+t) \le h^s(0)$ for each $x\in[0,H^s)$, and we obtain by the dominated convergence theorem that

$$\int_{[0,H^s)} \frac{g^s(x+\tau_{n^*}+t)}{\bar{G}^s(x)}\nu_0(dx) \le \int_{[0,H^s)} \frac{h^s(0)\bar{G}^s(x+\tau_{n^*}+t)}{\bar{G}^s(x)}\nu_0(dx) \to 0 \text{ as } t\to\infty.$$

The rest of the proof follows as in Case 1b. The last four displays imply that there exists $T\in(0,\infty)$ such that $B'(t) > \varepsilon/4$ whenever $B(t) < 1$ for a.e. $t\in[T,\infty)$. By the boundedness of $B$ it follows that there exists $T' > T$ such that $B(T') = 1$. Thus, for a.e. $t \ge T'$, we have $B'(t) > \varepsilon/4$ whenever $B(t) < 1$. In turn (by Lemma B.1) this implies that $B(t) = 1$ for all $t\in[T',\infty)$. Since all possible cases have been considered, this concludes the proof of the proposition. □

We now consider convergence properties of the measure-valued age process.

**Lemma 4.15.** *For $\lambda \ge 1$, under the assumptions of Proposition 4.12, suppose there exists $T<\infty$ such that $B(t) = 1$ for all $t \ge T$. Then $\nu_t \Rightarrow \nu_*$ and $\langle h^s, \nu_t\rangle \to 1$ as $t\to\infty$.*

*Proof.* By invoking Lemma 2.7, we can assume without loss of generality that $T = 0$. Then $B(t) = \langle \mathbf{1}, \nu_t\rangle = 1$ for all $t\ge 0$, and so by (4.6) of Corollary 4.4 of [25], $K$ has the representation

$$K(t) = \int_0^t \left(\int_{[0,H^s)} \frac{G^s(x+t-s) - G^s(x)}{\bar{G}^s(x)}\nu_0(dx)\right) dU_s(dx),\ t\ge 0.$$

In view of the representation for the fluid age measure in (2.19), the convergence $\nu_t \Rightarrow \nu_*$ is then a direct consequence of Lemma 6.2 of of [25] with $\pi = \nu$. Finally, since $h^s$ is bounded and monotone by Assumption 3.1(2), the set of its discontinuities is countable and thus has zero Lebesgue measure. Since $\nu_*$ is an absolutely continuous measure, the continuous mapping theorem implies $\langle h^s, \nu_t\rangle \to \langle h^s, \nu_*\rangle = \int_0^\infty g^s(x)dx = 1$, as $t\to\infty$. This concludes the proof of the lemma. □

4.4. **Uniqueness of Random Fixed Points.** We now show how the convergence results of the last two sections can be bootstrapped to conclude, under Assumption 2.8, the existence of a unique random fixed point.

**Proposition 4.16.** *Suppose $\lambda\ge 1$, Assumptions 2.1 and 2.8 hold and suppose that for any solution $(X,\nu,\eta)$ to the fluid equations with arrival rate $\lambda$ and initial condition $(X(0),\nu_0,\eta_0)\in\mathfrak{S}$,*

$$\eta_t \Rightarrow \lambda\eta_* \quad \text{and} \quad B_t \to 1. \tag{4.56}$$

*Then any random fixed point $\mu$ for the fluid equations with arrival rate $\lambda$ satisfies $\mu = \delta_{z_*^\lambda}$, where $z_*^\lambda = (x_*^\lambda, \nu_*, \lambda\eta_*)$, with $x_*^\lambda$ being the unique element of $\mathcal{X}_\lambda$ in (2.24).*

*Proof.* Fix $\lambda \geq 1$ and let $\mu$ be a random fixed point for the fluid equations with arrival rate $\lambda$. Let $(X(0), \nu_0, \eta_0)$ be a random element taking values in $\mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r)$ with law $\mu$ and let $(X, \nu, \eta)$ be the solution to the fluid equations with arrival rate $\lambda$ and initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$. Since $\eta_t \Rightarrow \lambda\eta_*$ and $B_t \to 1$ by assumption and the laws of $\eta_t$ and $\nu_t$ are invariant in $t$ since $\mu$ is a random fixed point, we have $\mathbb{P}(\eta_0 = \lambda\eta_*) = 1$ and $\mathbb{P}(B_t = \langle \mathbf{1}, \nu_t\rangle = 1) = 1$. Further, by continuity of $B$, we have $\mathbb{P}$-almost surely, $B_t = 1$ for all $t \geq 0$. Then by Lemma 4.15, it follows that $\nu_t \Rightarrow \nu_*$ as $t \to \infty$. Since the law of $\nu_t$ is invariant in $t$, it follows that $\mathbb{P}(\nu_0 = \nu_*) = 1$.

To complete the proof, it only remains to show that $\mathbb{P}(X(0) = x_*^\lambda) = 1$. Since almost surely, for all $t \geq 0$, $B(t) = 1$ and $\eta_t = \lambda\eta_*$, the relations (2.13) and (2.12) show that almost surely for all $t \geq 0$, $X(t) = Q(t) + 1$ and

$$R(t) = \int_0^t \left( \int_0^{Q(s)} h^r(\left(F^{\lambda\eta_*}\right)^{-1}(y))dy \right) ds = \lambda \int_0^t G^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(s)))ds.$$

Moreover, using the fact that almost surely for each $t \geq 0$, $\nu_t = \nu_*$, and hence, $D(t) = t\langle h^s, \nu_*\rangle = t$, we have from (2.13), (2.11) and the fact that $E = E^\lambda$ that almost surely for each $t \geq 0$,

$$Q(t) = Q(0) + (\lambda - 1)t - \lambda \int_0^t G^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(s)))ds \tag{4.57}$$

$$= Q(0) + \int_0^t \left( \lambda \bar{G}^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(s))) - 1 \right) ds. \tag{4.58}$$

We now consider two cases.
*Case 1:* $\lambda = 1$. In this case, we have

$$\int_0^t \left( \lambda \bar{G}^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(s))) - 1 \right) ds = -\int_0^t G^r((F^{\eta_*})^{-1}(Q(s)))ds.$$

It is clear from (4.58) that $Q$ is decreasing on $[0, \infty)$. By the non-negativity of $Q$, $q_* := \lim_{t\to\infty} Q(t)$ exists and the fact that $X(t) = Q(t) + 1$ implies $\lim_{t\to\infty} X(t) = x_* := q_* + 1$. Note that $G^r((F^{\eta_*})^{-1}(q_*)) = 0$ since, otherwise, $Q(t) \to -\infty$ as $t \to \infty$, which contradicts the non-negativity of $Q$. Therefore, by Assumption 2.8, the definition of $\mathcal{X}_\lambda$ in (2.24) and the fact that $\lambda - 1 = 0$, it follows that $x_*$ is equal to the unique element $x_*^\lambda$ of $\mathcal{X}_\lambda$. As before, since $\mu$ is a random fixed point, this implies that $\mathbb{P}(X(0) = x_*^\lambda) = 1$.
*Case 2:* $\lambda > 1$. In this case, it is clear from (4.58) that $Q$ is differentiable on $(0, \infty)$, and

$$Q'(t) = \lambda \bar{G}^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(s))) - 1 \text{ for each } t > 0. \tag{4.59}$$

First note that since by (2.14), $Q(t) \leq \langle \mathbf{1}, \eta_t\rangle = \lambda\langle \mathbf{1}, \eta_*\rangle$, $Q(t)$ is bounded. We now argue by contradiction to show that $q_* = \lim_{t\to\infty} Q(t)$ exists. Suppose this is not the case. Then, since $Q$ is bounded on $[0, \infty)$, $Q$ must oscillate for the limit not to exist. By the continuity of $Q$, this implies there must exist two sequences of times $\{t_n, n \geq 1\}$ and $\{s_n, n \geq 1\}$ with $t_n \to \infty$, $s_n \to \infty$ as $n \to \infty$, and $\varepsilon > 0$ such that $|Q(t_n) - Q(s_n)| > \varepsilon$ for all $n$ sufficiently large, and $Q'(t_n) = Q'(s_n) = 0$ for each $n \geq 1$. By (4.58), the latter relation implies

$$G^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(t_n))) = G^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(s_n))) = \frac{\lambda - 1}{\lambda} \text{ for all } n \geq 1.$$

Since $Q$ is bounded, there exist $0 \leq q_i \leq \lambda\langle \mathbf{1}, \eta^r\rangle$, $i = 1, 2$ and is a subsequence $\{n_k, k \geq 1\}$ such that $Q(t_{n_k}) \to q_1$ and $Q(s_{n_k}) \to q_2$ as $k \to \infty$. It follows that $|q_1 - q_2| \geq \varepsilon$ and $G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_1)) = G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_2)) = \frac{\lambda-1}{\lambda}$, where we have used the fact that $G^r$ and $\left(F^{\lambda\eta_*}\right)^{-1}$ are continuous, with the latter continuity holding because $\lambda\eta_*$ has a density $\lambda\bar{G}^r$ that is strictly positive on its support. By Assumption 2.8, we have $q_1 = q_2$ which contradicts $|q_1 - q_2| \geq \varepsilon$. Thus, $q_* = \lim_{t\to\infty} Q(t)$ exists and then $\lambda\bar{G}^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(t))) - 1 = 0$ since otherwise by (4.58) $Q$ will not have a limit. We can then argue as in Case 1 that $X(t) \to q_* + 1 = x_*^\lambda$, and thus $\mathbb{P}(X(0) = x_*^\lambda) = 1$. This completes the proof of the theorem. □

## 5. Results Regarding the Multiclass Model

Here we consider the model with multiple classes operating under a fixed priority discipline. This model, with general class-dependent service time and patience time distributions, was analyzed in [9] and convergence at the fluid scale, uniformly on compact time intervals, was established. Here, we study the long-time behavior under the additional assumption that the service time distribution does not depend on the job class, and its hazard rate function is bounded away from zero and infinity, that is, satisfies Assumption 3.1(1). For simplicity of exposition, we also assume that the reneging distributions are exponential (but may depend on the job class) since the main motivation is to deduce the optimality of a certain priority scheduling rule (known as the $c\mu/\theta$ rule; see details below) discussed in [9], which is not expected to hold beyond the exponential reneging case. As shown in [9], this optimality result relies on the convergence of the invariant distributions of the fluid-scaled process, as $N \to \infty$, to the unique element of the invariant manifold of the fluid limit (under assumptions that ensure such uniqueness). However, the convergence result in [9] (specifically Theorem 4.3 therein) suffers from the same flaw as that described for the single-class case in Remark 3.3; namely, from the proof in [9] one can only deduce that the invariant distributions of the $N$-server systems exist, are tight and that any subsequential limit of the sequence of invariant distributions must be a *random* fixed point of the fluid equations (defined analogously to Definition 2.10). As explained in Remark 2.11 in the single-class setting, in order to show that there is a unique random fixed point (which must then coincide with the unique element of the invariant manifold) it suffices to establish the long-time convergence of the solution of the fluid equations with any initial condition to the unique element of the invariant manifold. Thus, the limit interchange result that we prove here fixes the flaw in the main optimality result of [9] under the additional assumptions on the service distribution stated above. This leaves open the question of whether there is also a limit interchange for class dependent service times and when hazard rates are not necessarily bounded. We present the fluid equations in Section 5.1, and then state and prove the theorem in Section 5.2.

5.1. **Fluid Model Equations for the Multiclass System.** Analogous to the single-class case, for each class $i \in \{1, \ldots, J\}$, we denote by $B_i$, $X_i$ and $Q_i$ nonnegative functions that represent the fluid analogs of the number in service, number in system and number in queue, let the nonnegative, nondecreasing functions $D_i$, $K_i$ and $R_i$ represent the fluid analogs of cumulative class $i$ departures from service, cumulative entries to service and cumulative reneging, and let $\nu_i$ represent the fluid analog of the measure-valued function that encodes the ages of class $i$ jobs in service. Since we assume exponential reneging times, we will not require the potential reneging measures $\eta_i$, but only the reneging rate $\theta_i > 0$. Also, let $(X, \theta, \nu, B, Q, D, K, R)$ be the corresponding vector-valued processes whose $i$th component is given by $(X_i, \theta_i, \nu_i, B_i, Q_i, D_i, K_i, R_i)$. We describe the fluid equations only for the special case when all service distributions are identical, with common cumulative distribution function $G = G^s$, hazard rate function $h = h^s$ and support $[0, H) = [0, H^s) = [0, \infty)$, and arrival rates $\lambda_i > 0, i = 1, \ldots, J$.

Before we present the fluid model equations, let us comment on the special form that the single-server fluid model equations (of Definition 2.3) take when the reneging is exponential. In this case, the reneging hazard rate is constant, namely $h^r(t) = \theta$ for all $t$, and thus equation (2.12) takes the form

$$R(t) = \theta \int_0^t Q(s)ds,$$

and thus there is no longer any need to keep track of the potential reneging measure $\eta$. Accordingly, in the multiclass setting, our fluid model is an extension of such a modified set of fluid equations where the equation of $R$ is similar to the above display, and from which $\eta$ is absent.

**Definition 5.1.** *Given arrival and reneging rate vectors $\lambda \in (0, \infty)^J$ and $\theta \in (0, \infty)^J$, and initial condition $(X(0), \nu_0) \in [0, \infty)^J \times (\mathcal{M}_F[0, \infty))^J$, a tuple $(B, X, Q, D, K, R, \nu) \in (\mathcal{D}_{\mathbb{R}_+^J}(\mathbb{R}_+))^3 \times (\mathcal{D}^+_{\mathbb{R}_+^J}(\mathbb{R}_+))^3 \times (\mathcal{D}_{\mathcal{M}_F[0,\infty)}(\mathbb{R}_+))^J$ is said to be a solution to the multiclass fluid equations with initial condition $(X(0), \nu_0)$ and arrival and reneging rate vectors $\lambda$ and $\theta$ if equations (5.1)–(5.2) below are satisfied: For $\varphi \in \mathcal{C}^1_c([0, \infty) \times \mathbb{R}_+)$,*

*and* $t \geq 0$,

$$\langle \varphi(\cdot,t), \nu_{i,t}\rangle = \langle \varphi(\cdot,0), \nu_{i,0}\rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_t(\cdot,s), \nu_{i,s}\rangle ds - \int_0^t \langle h(\cdot)\varphi(\cdot,s), \nu_{i,s}\rangle ds + \int_0^t \varphi(0,s) dK_i(s), \tag{5.1}$$

*where* $B, D, R$ *are the auxiliary processes given by*

$$B_i(t) = \langle 1, \nu_{i,t}\rangle, \qquad D_i(t) = \int_0^t \langle h, \nu_{i,s}\rangle ds, \qquad R_i(t) = \theta_i \int_0^t Q_i(s) ds. \tag{5.2}$$

*and for* $t \geq 0$, $K, B, D$ *satisfy the following balance equations and basic relations:*

$$B_i = B_{i,0} - D_i + K_i, \tag{5.3}$$

$$X_i(t) = X_{i,0} - D_i(t) + \lambda_i t - R_i(t), \tag{5.4}$$

$$Q_i = X_i - B_i, \tag{5.5}$$

*as well as conditions imposing work conservation and non-preemptive priority:*

$$I := 1 - \sum_{i=1}^J B_i = \Big(1 - \sum_{i=1}^J X_i\Big)^+, \tag{5.6}$$

$$K_i(t) = \int_{[0,t]} 1_{\{\sum_{j=1}^{i-1} Q_{j,s}=0\}} dK_i(s), \qquad i \geq 2,\ t \geq 0. \tag{5.7}$$

Under the assumption of bounded reneging hazard rates, which is indeed fulfilled when the reneging distribution is exponential, it was shown in [9, Theorem 3.1] that uniqueness holds for solutions of the fluid equations for any given data and initial conditions. Existence of solutions was also established there by showing that the scaling limit of the underlying queueing system is a solution.

By the same argument given in the proof of Theorem 2.6, it follows from the results in Theorem 4.1 of [25] that the measure-valued age equation (5.1) implies that for every $\psi \in \mathcal{C}_b([0,\infty))$ or $\psi = h$,

$$\langle \psi, \nu_{i,t}\rangle = \int_{[0,\infty)} \frac{\bar{G}(x+t)}{\bar{G}(x)} \psi(x+t)\nu_{i,0}(dx) + \int_{[0,t]} \bar{G}(s-x)\psi(s-x) dK_i(s), \tag{5.8}$$

where recall $\bar{G} = 1 - G$. In what follows, given a vector or vector-valued process $Y$, we use $\tilde{Y}$ to be generic notation for the sum $\sum_{i=1}^J Y_i$. By (5.8), $\tilde{\nu}$ and $\tilde{K}$ satisfy, for every $\psi \in \mathcal{C}_b([0,\infty))$ or $\psi = h$,

$$\langle \psi, \tilde{\nu}_t\rangle = \int_{[0,\infty)} \frac{\bar{G}(x+t)}{\bar{G}(x)} \psi(x+t)\tilde{\nu}_0(dx) + \int_{[0,t]} \bar{G}(t-s)\psi(t-s) d\tilde{K}(s). \tag{5.9}$$

In other words, (2.19) holds with $(\nu, K)$ and $G^s$ replaced with $(\tilde{\nu}, \tilde{K})$ and $G$. We now argue that, $\tilde{K}$ and $\tilde{B}$ satisfy the analog of (2.20). First, note that by (5.2), $\tilde{B} = \langle 1, \tilde{\nu}\rangle$, and if $\tilde{B}_t < 1$ then, on an open interval containing $t$ we have $\tilde{X} < 1$ due to (5.6). Hence, $\tilde{Q} = 0$ by (5.5) and $\tilde{R} = 0$ by (5.2). Hence, subtracting (5.4) from (5.3), $\tilde{K} = \tilde{E} + c$ on this interval (where $c$ does not depend on time), and so $\tilde{K}'(s) = \tilde{\lambda}$ holds on the interval. Combining this with

$$\tilde{K}(t) = \tilde{B}(t) - \tilde{B}(0) + \int_0^t \langle h, \tilde{\nu}_s\rangle ds,$$

which follows from (5.3) and (5.2), we obtain, exactly as in [9, Theorem 3.2], that for a.e. $t$, $\tilde{K}'(t) = \tilde{k}(t)$ where

$$\tilde{k}(t) = \begin{cases} \langle h, \tilde{\nu}_t\rangle, & \tilde{B}(t) = 1, \\ \tilde{\lambda}, & \tilde{B}(t) < 1. \end{cases} \tag{5.10}$$

5.2. **Results for the Multiclass System.** We will be interested in the supercritical case where $\sum_i \lambda_i > 1$ and $\theta_{\min} = \min_i \theta_i > 0$. Let $\rho_i$, $i = 1, \ldots, J$ be characterized by

$$\sum_{i=1}^{j} \rho_i = \Big(\sum_{i=1}^{j} \lambda_i\Big) \wedge 1, \qquad j = 1, \ldots, J,$$

and let

$$q_i = \frac{\lambda_i - \rho_i}{\theta_i}, \qquad i = 1, \ldots, J.$$

We now state the main result.

**Theorem 5.2.** *Suppose that $h$ satisfies Assumption 3.1, and $\lambda, \theta \in (0,\infty)^J$ are such that $\tilde{\lambda} = \sum_{i=1}^J \lambda_i > 1$, and $(X_0, \nu_0) \in \mathbb{R}_+^J \times (\mathcal{M}_F[0,\infty))^J$ satisfies $1 - \langle \mathbf{1}, \tilde{\nu}\rangle = (1 - \tilde{X})^+$. Then any solution $(B, X, Q, D, K, R, \nu)$ to the multiclass fluid equations with initial condition $(X_0, \nu_0)$ and arrival and reneging rate vectors $\lambda$ and $\theta$ satisfies $\nu_{i,t} \Rightarrow \rho_i \nu^*$ and $Q_i(t) \to q_i$ as $t \to \infty$ for $i = 1, \ldots, J$.*

**Remark 5.3.** This validates Theorem 5.1 of [9] in the special case where for all $i$, $h_i^s = h$, with $h$ satisfying Assumption 3.1.

**Remark 5.4.** The characterizations in (5.8) and (5.10) show that the aggregate processes $(\tilde{X}, \tilde{\nu})$ and $(\tilde{D}, \tilde{K}, \tilde{R}, \tilde{S}, \tilde{Q}, \tilde{B})$ satisfy the fluid equations of the single class case (see Definition 2.3), subject to the simplification described at the beginning of Section 5.1, where in particular reneging is given directly by (5.2) and the process $\eta$ is not used. Hence, in the supercritical setting $\tilde{\lambda} > 1$, we may conclude from Theorem 3.2(2) that, with $\nu^*(dx) = \bar{G}(x)dx$, one has $\tilde{\nu}_t \Rightarrow \nu^*$ and that there exists $T < \infty$ such that $\tilde{B}_t = 1$ for all $t \geq T$. Moreover, by Proposition 4.8, $\langle h, \tilde{\nu}_t\rangle \to 1$. As a result, by (2.20), one has $\tilde{k}_t = \langle h, \tilde{\nu}_t\rangle$ for all large $t$, and hence also $\tilde{k}_t \to 1$.

*Proof of Theorem 5.2.* In this proof, the special case in which there exists $i_0 \in \{1, \ldots, J-1\}$ such that $\sum_{i=1}^{i_0} \lambda_i = 1$ is called the *borderline case*, and the more typical case, where such $i_0$ does not exist, is called the *typical case*.

If $\lambda_1 < 1$, set $\ell := \max\{j : \sum_{i=1}^j \lambda_i < 1\}$, otherwise let $\ell = 0$. Also, set $m = \ell + 1$. Then, since by assumption $\tilde{\lambda} > 1$, by the definition of $m$, we have $\sum_{i=1}^m \lambda_i = 1$ [respectively, $> 1$] in the borderline case [respectively, in the typical case]. Also, in what follows, we use the hat (when $\ell \geq 1$) and # notation for summation up to $l$ and, respectively, $m$, as in

$$\hat{Y} = \sum_{i=1}^{\ell} Y_i, \quad \text{and} \quad Y^{\#} = \sum_{i=1}^{m} Y_i, \qquad Y = \lambda, X, \nu, D, K, R, B. \tag{5.11}$$

(in addition to the notation already introduced, $\tilde{Y} = \sum_{i=1}^J Y_i$).

The structure of the proof is as follows. In Step 1 we prove the assertions for $i \leq \ell$. Steps 2 and 3 address the remaining classes $i \geq m$ in the typical and borderline cases, respectively. First, note that since $\tilde{\lambda} > 1$, by Remark 5.4, there exists $T < \infty$ such that

$$\tilde{k}(t) \to 1, \langle h, \tilde{\nu}_t\rangle \to 1 \text{ as } t \to \infty \quad \text{and} \quad \tilde{B}(t) = 1 \text{ for all } t \geq T. \tag{5.12}$$

**Step 1.** Consider the case $\ell \geq 1$ (that is, $\lambda_1 < 1$). In this step we consider classes $1 \leq i \leq \ell$ and establish the claim that there exists $t_1 < \infty$ such that $Q_i(t) = 0$ for all $t \geq t_1$, and moreover, that $\nu_i(t) \to \rho_i \nu^*$ as $t \to \infty$. (Note that for $i \leq \ell$, $\lambda_i = \rho_i$, hence the asserted convergence $Q_i(t) \to q_i = 0$ would then follow).

Recalling the notational convention (5.11), by the definition of $\ell$, $\hat{\lambda} = \sum_{i=1}^{\ell} \lambda_i < 1$, and so there exist $\varepsilon_0 > 0$ and $0 < t_0 < \infty$ such that $\langle h, \tilde{\nu}_t\rangle > \hat{\lambda} + \varepsilon_0$ for all $t \geq t_0$. If $\hat{Q}(t) = 0$ for all $t \geq t_0$ then the claim follows trivially. So, we now consider the converse case, when $\mathcal{O} := \{t > t_0 : \hat{Q}(t) > 0\}$ is non-empty. Since $\hat{Q}$ is continuous, $\mathcal{O}$ is open and is a union of countable open intervals. For a.e. $s$ in each such interval, by

(5.7), for all $i > \ell$, $K_i'(s) = 0$. Moreover, since (5.5) and (5.6) together show that, $\tilde{Q}(t) > 0$ implies $\tilde{B}(t) = 1$ for any $t > 0$, we conclude in particular that $\tilde{B}(s) = 1$. In turn, (5.3), (5.5), (5.2) and the fact that $\tilde{R}$ is non-decreasing together imply that for a.e. $s \in \mathcal{O}$, $\tilde{D}'(s) = \tilde{K}'(s) = \langle h, \tilde{\nu}_s \rangle$. Thus, we have for a.e. $s > t_0$,

$$\hat{Q}(s) > 0 \quad \Rightarrow \quad \hat{Q}'(s) = \hat{\lambda} - \hat{R}'(s) - \hat{K}'(s) \leq \hat{\lambda} - \langle h, \tilde{\nu}_s \rangle \leq -\varepsilon_0.$$

Thus, there must exist a finite time, $t_1 \geq t_0$, when $\hat{Q}(t_1) = 0$. Since the last display continues to hold for all $s \geq t_1$, applying Lemma B.1 with $f = -Q$, $T = t_1$, $S = \infty$ and $c = 0$, it follows that for all $t \geq t_1$, $\hat{Q}(t) = 0$, or equivalently, $Q_i(t) = 0$ for all $i \leq \ell$.

To finish proving the claim in Step 1, it only remains to show that $\nu_i(t) \to \rho_i \nu^* = \lambda_i \nu^*$ for $i \leq \ell$. For $t \geq t_1$, it follows from (5.5) and (5.2), respectively, that for $i \leq \ell$, $X_i(t) = B_i(t)$ and $R_i'(t) = 0$. Hence by (5.3)–(5.4), $K_i'(t) = \lambda_i$ for such $i$ and $t$. Substituting these relations in (5.8) and taking the large $t$ limit yields (exactly as in the proof of Lemma 4.1), the convergence of $\nu_i(dx)$ to $\lambda_i \bar{G}(x)dx$ as asserted.

**Step 2.** In this step we treat the typical case, proving the claim for all the remaining classes $i \geq m = \ell + 1$ (where possibly $\ell = 0$, $m = 1$). Recall the notation in (5.11) and note that in this case one has $\lambda^\# = \sum_{i=1}^m \lambda_i > 1$. By the definition of $m$ and $\rho_i$, this implies $\rho_m < \lambda_m$.

Let $t_1 < \infty$ be as in Step 1, and assume without loss of generality that $t_1 \geq T$, where $T$ is as in (5.12). Then given $\varepsilon \in (0, 1 - \lambda^\#)$, there exists $t_2 = t_2(\varepsilon) \geq t_1$ such that for all $t \geq t_2$, $|\langle h, \tilde{\nu}_t \rangle - 1| < \varepsilon$, $|\tilde{k}(t) - 1| < \varepsilon$ and $\tilde{B}(t) = 1$. Then (5.2) implies that $\tilde{D}_t'(t) = \langle h, \tilde{\nu}_t \rangle \leq (1 + \varepsilon)$, and since clearly, $(D^\#)' \leq \tilde{D}'$, on $[t_2, \infty)$, we have for all $\varepsilon_1 \leq \lambda^\# - 1 - \varepsilon$,

$$\frac{dX^\#}{dt} = \lambda^\# - \frac{dD^\#}{dt} - \frac{dR^\#}{dt} \geq \lambda^\# - (1 + \varepsilon) - \theta_m Q_m \geq \varepsilon_1 - \theta_m Q_m.$$

We now argue by contradiction to prove the claim that there exists $t_3 \geq t_2$ such that $Q_m(t_3) > 0$. Indeed, assume $Q_m$ vanishes on the whole interval $[t_2, \infty)$. Then the last display shows that $X^\#(t) \to \infty$, and hence by (5.5) and the fact that (5.6) implies $\tilde{B}$ lies in $[0, 1]$, $Q^\#(t) \to \infty$. But since $\hat{Q}$ vanishes on $(t_1, \infty) \supset (t_2, \infty)$ by Step 1, this implies $Q_m(t) = Q^\#(t) - \hat{Q}(t) = Q^\#(t) \to \infty$, which contradicts the assumption that $Q_m$ is identically zero on $[t_2, \infty)$. This proves the claim.

Let $t_3 \geq t_2$ be such that $Q_m(t_3) > 0$, and let $\mathcal{O}_m := \{s \in [t_3, \infty) : Q_m(s) > 0\}$. We show below that $\mathcal{O}_m = [t_3, \infty)$. Towards this goal, we will find it more convenient to work with the balance equation for $Q^\#$ than with $X^\#$. That is, using (5.3)–(5.5) and (5.2), note that

$$Q_i' = \lambda_i - K_i' - \theta_i Q_i, \quad i = 1, \ldots, J, \qquad \text{and} \qquad \frac{dQ^\#}{dt} = \lambda^\# - \frac{dK^\#}{dt} - \sum_{i=1}^m \theta_i Q_i^\#. \tag{5.13}$$

On any open interval in $\mathcal{O}_m$, $Q_m > 0$ and $\hat{Q} = 0$, and hence the priority rule (5.7) implies $dK^\#/dt = d\tilde{K}/dt = \tilde{k}$, where recall $|\tilde{k}(t) - 1| < \varepsilon$. Thus, for all $t \geq t_3$, we have

$$Q_m(t) > 0 \quad \Rightarrow \quad Q_m'(t) = \frac{dQ^\#}{dt}(t) \geq \lambda^\# - (1 + \varepsilon) - \frac{dR^\#}{dt}(t) \geq \varepsilon_1 - \theta_m Q_m(t).$$

Since this is strictly greater than $\varepsilon_1/2$ whenever $Q_m(t) < \varepsilon_1/2\theta_m$, this clearly implies $Q_m(t) > 0$ for all $t \in [t_3, \infty)$, as claimed. In turn, by the priority rule (5.7), this implies that on $[t_3, \infty)$, $K_i' = 0$ for all $i > m$, and therefore by (5.3) and (5.2), $B_i' = -\langle h, \nu_i \rangle \leq -\varepsilon_h B_i$, where recall that $\varepsilon_h$ is the strictly positive lower bound on $h$. This shows that for $i > m$, $B_i(t) \to 0$ as $t \to \infty$ and hence, $\nu_i \Rightarrow 0$. As $t \to \infty$, since we already have convergence of the aggregate $\tilde{\nu}_t \Rightarrow \nu^*$ (see Remark 5.4) and $\nu_{i,t} \Rightarrow \lambda_i \nu^*$ for all $i < m$ (by Step 1), we conclude that $\nu_{m,t} \Rightarrow \rho_m \nu^*$.

To complete Step 2, it only remains to address the convergence of $Q_i$, $i \geq m$. Since, as argued above, for $t \in [t_3, \infty)$, $K_i'(t) = 0$ for $i > m$, (5.13) shows that $Q_i(t) \to \lambda_i/\theta_i = q_i$ as $t \to \infty$. As for $Q_m$, note that since on $[t_3, \infty)$, for $1 \leq i \leq \ell = m - 1$, $Q_i = 0$ by Step 1, (5.13) shows that $K_i' = \lambda_i$, or equivalently, $\hat{K}' = \hat{\lambda}$.

Thus, denoting $e(t) := \tilde{k}(t) - 1$, we have $e(t) \to 0$, and recalling that $\rho_m = \hat{\lambda} - 1$,

$$K_m'(t) = \tilde{K}'(t) - \hat{K}_i'(t) = \tilde{k}(t) - \hat{\lambda} = (\rho_m + e(t)).$$

Thus, we obtain

$$Q_m'(t) = (\lambda_m - \rho_m - e(t)) - \theta_m Q_m(t).$$

This implies that as $t \to \infty$, $Q_m(t)$ converges to $q_m = (\lambda_m - \rho_m)/\theta_m$. Here, we used the elementary fact that for a differentiable function $u$ on $[0, \infty)$,

$$u'(t) = w(t) - \theta u(t) \text{ and } u(0) = u_0 \qquad \Rightarrow \qquad u(t) = \int_0^t e^{-\theta(t-s)} w(s) ds + u_0 e^{-\theta t}, \tag{5.14}$$

which converges to $c/\theta$ whenever $w(t) \to c$ as $t \to \infty$.

**Step 3.** Lastly, we consider the borderline case, and establish the assertions regarding the remaining classes $i \in \{m, \ldots, J\}$. In this case $\lambda^\# = \sum_{i=1}^m \lambda_i = 1$.

As in (5.10), the priority structure specified by (5.7) dictates that $dK^\#/dt = k^\#$, where $k^\#(t)$ is given by $\langle h, \tilde{\nu}_t \rangle$ when $B^\#(t) = 1$ and equal to $\lambda^\#$ when $B^\#(t) < 1$. Since $\lambda^\# = 1$ and by (5.12), $\langle h, \tilde{\nu}_t \rangle \to 1$ we infer that $k^\#(t) \to 1$ as $t \to \infty$. Summing (5.8) over $i \leq m$, using $\int_0^\infty \bar{G}(x) dx = 1$ and applying the test function $\psi = \mathbf{1}$ shows that $B_t^\# \to 1$ as $t \to \infty$ (where the application of bounded continuous $\psi$ can be justified in the usual manner). Applying general compactly supported test functions gives $\nu_t^\# \Rightarrow \nu^*$, where $\nu^*(dx) = \bar{G}(x) dx$. Given the convergence already established for $\nu_{i,t}$, $i \leq \ell = m - 1$, the convergence of $\nu_t^\#$ yields that of $\nu_{m,t} \to \lambda_m \nu^*$ (note that in the borderline case currently considered, $\lambda_m = \rho_m$). Moreover, the fact that $B^\#(t) \to 1$ implies that $\sum_{i=m+1}^J B_i(t) = \tilde{B}(t) - B^\#(t) \to 0$, and hence, for all $i > m$, $B_i(t) \to 0$ and consequently $\nu_{i,t} \Rightarrow 0$.

Next we show that $Q_i(t) \to q_i$ for $i > m$, for which we again use (5.13). Combining the convergence $k^\#(t) \to 1$ that we just showed with $\tilde{k}(t) \to 1$ from (5.12), it follows that $k_i(t) \to 0$ for all $i > m$. Recalling that $K_i' = k_i$ and using the first equation in (5.13) and (5.14) yields $Q_i(t) \to \lambda_i/\theta_i = q_i$, for $i > m$.

We finally show that $Q_m(t) \to 0$. To this end, note that by the aggregate equation in (5.13) and the property that for sufficiently large $t$, $Q_i(t) = 0$, for $i \leq \ell$, (from Step 1) giving $dR^\#(t)/dt = \theta_m Q_m(t) = \theta_m Q^\#(t)$. Since $\lambda^\# = 1$, (5.13) shows that the following is valid for all large $t$,

$$\frac{dQ^\#}{dt} = 1 - k^\#(t) - \theta_m Q^\#(t).$$

Recalling that $k^\#(t) \to 1$, and again using (5.14), it follows that $Q^\#(t) \to 0$. Consequently, $Q_m(t) \to 0$. This completes the proof. □

## Appendix A. Proof of Lemma 4.5.

*Proof of Lemma 4.5.* Fix a measurable function $f : [0, \infty) \mapsto \mathbb{R}_+$ with $\int_0^\infty f dx \leq 1$. For notational conciseness, define

$$Af := \int_0^\infty h^s(x) f(x) \log \frac{f(x)}{f^*(x)} dx - z_f \log z_f,$$

where recall $z_f := \int_0^\infty h^s f dx < \infty$. Let $U(x) := x \log x$, $x > 0$, $U(0) = 0$. Fix a non-negative measurable function $\psi$ on $[0, \infty)$ with $c_\psi = \int_0^\infty \psi dx \leq 1$. Then, note from the definition of $A$ in (4.8) that

$$A(\psi) = \int_0^\infty U\left(\frac{\psi}{f^*}\right) h^s f^* dx - U(z_\psi). \tag{A.1}$$

Since $\int_0^\infty h^s f^* dx = 1$ and $\int_0^\infty \left(\frac{\psi}{f^*}\right) h^s f^* dx = z_\psi$, the convexity of $U$ and Jensen's inequality imply the nonnegativity of $A(\psi)$. To obtain the more refined estimate (4.8), define

$$V(x) := U(x) - [U'(z_f)(x - z_f) + U(z_f)].$$

Then, the strict convexity of $U$ implies $V(x) \geq 0$ and $V(x) = 0$ if and only if $x = z_f$. Using (A.1) we have

$$A(\psi) = \int_0^\infty V\Big(\frac{\psi}{f^*}\Big) h^s f^* dx + \int_0^\infty U'(z_f)\Big(\frac{\psi}{f^*} - z_f\Big)\Big] h^s f^* dx = \int_0^\infty V\Big(\frac{\psi}{f^*}\Big) h^s f^* dx,$$

where the last equality uses the definition of $z_\psi$. Since $V \geq 0$, denoting $c_\psi := \int_0^\infty \psi dx \leq 1$, and recalling the functional $R$ from (4.5), we have

$$\begin{aligned} A(\psi) &\geq \varepsilon_h \int_0^\infty V\Big(\frac{\psi}{f^*}\Big) f^* dx = \varepsilon_h \Big[\int_0^\infty \psi \log \frac{\psi}{f^*} dx - \int_0^\infty U'(z_\psi)\Big(\frac{\psi}{f^*} - z_\psi\Big) f^* dx - U(z_\psi)\Big] \\ &= \varepsilon_h [R(\mu^\psi \| \nu_*) - c_\psi U'(z_\psi) + U'(z_\psi) z_\psi - U(z_\psi)] \\ &= \varepsilon_h [R(\mu^\psi \| \nu_*) - c_\psi \log z_\psi - c_\psi + z_\psi] \\ &= \varepsilon_h \{R(\mu^\psi \| \nu_*) + c_\psi[-\log z_\psi - 1 + z_\psi] + z_\psi(1 - c_\psi)\} \\ &\geq \varepsilon_h R(\mu^\psi \| \nu_*), \end{aligned}$$

where the third equality used the fact that $U'(x) = \log x + 1$ and $U'(x)x - U(x) = x$, and the last inequality uses the elementary inequality $x - \log x \geq 1$ for all $x > 0$. This proves (4.8). □

## Appendix B. An Elementary Property of Absolutely Continuous Functions

The following simple property is used in Section 4.3.

**Lemma B.1.** *Let $f$ be an absolutely continuous function defined on $[0, S)$ for some $0 < S \leq \infty$. Suppose that there exist a time $T \in (0, S)$ and a constant $c > 0$ such that $f(T) \geq c$ and for a.e. $t \in (T, S)$, $f'(t) \geq 0$ if $f(t) < c$. Then $f(t) \geq c$ for all $t \in [T, S)$.*

*Proof.* Suppose the conclusion of the lemma does not hold. Then there must exist $T < t_1 < t_2$ for which $f(t_1) \geq c$ and $f(t_2) < c$. Since $f$ is absolutely continuous, there must exist some interval $(s_1, s_2) \subset (t_1, t_2)$ such that $f(s) < c$ for $s \in (s_1, s_2)$ and $f'(s) < 0$ for $s$ in a subset $\mathcal{S} \subset (s_1, s_2)$ of positive Lebesgue measure. However, this contradicts the assumption of the lemma, that is, for a.e. $t \in [T, \infty)$, $f'(t) \geq 0$ if $f(t) < c$. Hence the lemma is proved. □

**Acknowledgement.** We would like to thank Amber Puha for raising the question of a full justification of the convergence results in Theorem 3.3 of [24].

## References

[1] Agarwal, P. and Ramanan, K. Invariant states of hydrodynamic limits of randomized load balancing networks. (2020) Preprint, https://arxiv.org/abs/2008.08510.
[2] Aghajani, R. and Ramanan, K. "Hydrodynamic limits of randomized load balancing networks," *Annals of Applied Probability*, 29, no. 4 (2019) 2114-2174.
[3] Aghajani, R. and Ramanan, K. Ergodicity of an SPDE associated with a many-server queue. *Annals of Applied Probability*, 29 (2019) no. 2, 994-1045.
[4] Aghajani, R. and Ramanan, K. The limit of stationary distributions of many-server queues in the Halfin-Whitt regime, *Mathematics of Operations Research*, 45 (2020) no. 3, 10161055.
[5] Aghajani, R., Li, X. and Ramanan, K. "The PDE method for the analysis of randomized load balancing networks", *Proc. ACM SIGMETRICS*, Meas. Anal. Comput. Syst. Vol. 1, No. 2 (2017), Article 38.
[6] Asmussen, S. Applied probability and queues, 2nd edition ed., (2003) Springer-Verlag, New York.
[7] Atar, R., Giat, C. and Shimkin, N. The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research* 58.5 (2010) 1427–1439.
[8] Atar, R., Giat, C. and Shimkin, N. On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems* 67, No. 2 (2011) 127–144.
[9] Atar, R., Kaspi, H. and Shimkin, N. Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research* 39, No. 3 (2013) 672-696.
[10] Baccelli, F. and Hebuterne, G. On queues with impatient customers. In Performance 81 (E. Gelenbe, ed.) (1981) 159179.

[11] Barlow, R., Marshall, A. and Proschan, F. Properties of Probability Distributions with Monotone Hazard Rate *The Annals of Mathematical Statistics* 34, No. 2 (1963) 375-389.

[12] Brown, M. Bounds, Inequalities, and Monotonicity Properties for Some Specialized Renewal Processes *Annals of Probability* 8, No. 2 (1980) 227-240.

[13] Caizo, J.A., Carrillo, J.A. and Cuadrado, S. Measure solutions for some models in population dynamics, *Acta Applicandae Mathematicae* 123 (2013) 141-156.

[14] Cinlar, E. *Introduction to Stochastic Processes*, (1975) Prentice-Hall.

[15] Csiszár, I and Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Cambridge University Press, 2011.

[16] Dai, J. G., Dieker, A.B. and Gao, X. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems* 78, No. 1 (2014) 1-29.

[17] Dai, J. G. and He, S., Many-Server Queues with Abandonment: A Survey of Diffusion and Fluid Approximations, *Journal of Systems Science and Systems Engineering*, 21, (2012) 136.

[18] Feller, W. *An Introduction to Probability Theory and its Applications*, 1957.

[19] Fu, Y. and Williams, R. J. Asymptotic Behavior of a Critical Fluid Model for Bandwidth Sharing with General File Size Distributions, *Preprint*, 2021.

[20] Gamarnik, D. and Goldberg, D.A. Steady-state $GI/G/n$ queue in the Halfin–Whitt regime. *The Annals of Applied Probability* 23, No. 6 (2013) 2382–2419.

[21] Gamarnik, D. and Stolyar, A.L. Multiclass multiserver queueing system in the Halfin–Whitt heavy traffic regime: Asymptotics of the stationary distribution. *Queueing Systems* 71, Nos. 1-2 (2012) 25-51.

[22] Garnett, O., Mandelbaum, A. and Reiman, M. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4, No. 3 (2002) 208-227.

[23] Kang, W. and Ramanan, K. Fluid limits of many-server queues with reneging. *The Annals of Applied Probability* 20 (2010) 2204–2260.

[24] Kang, W. and Ramanan, K. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *The Annals of Applied Probability*, 22, No. 2 (2012) 477-521.

[25] Kaspi, H. and Ramanan, K. Law of large numbers limits for many-server queues, *Annals of Applied Probability*, 21, No. 1 (2011) 33-114.

[26] Long, Z. and Zhang J. Virtual allocation policies for many-server queues with abandonment. *Mathematical Methods of Operations Research* 90 (2019), 399451.

[27] Long, Z., Shimkin, N., Zhang, H. and Zhang, J. Dynamic Scheduling of Multiclass Many-Server Queues with Abandonment: The Generalized $c\mu/\theta$ Rule. *Operations Research* (2020).

[28] Mandelbaum, A., Massey, W.A. and Reiman, M.I. Strong approximations for Markovian service networks. *Queueing Syst.* 30 (1998) 149201.

[29] Mandelbaum, A. and and Momčilović, P. Queues with Many Servers and Impatient Customers *Mathematics of Operations Research*, 37, No. 1 (2012), 41-65.

[30] Mandelbaum, A. and Zeltyn, S. Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers, *Operations Research*, 57, No. 5 (2009) 1189-1205.

[31] Mather, W.H., Cookson, N., Hasty, J., Tsimring, L.S. and Williams, R.J. Correlation resonance generated by coupled enzymatic processing. *Biophysical Journa*l, 99, No. 10 (2010) 3172 3181.

[32] Michel, P. , Mischler, S. and Perthame, B. General entropy equations for structured population models and scattering. *C.R. Acad. Sci. Paris, Ser. I*, 338 (2004) 697-702.

[33] Mulvany, J.A., Puha, A.L. and Williams, R.J. "Asymptotic Behavior of a Critical Fluid Model for a Multiclass Processor Sharing Queue via Relative Entropy." *Queueing Systems*, 93 (2019), 351-397.

[34] Niculescu, C.P. and Persson, L.-E. *Convex Functions and their Applications: A Contemporary Approach* (2005), Springer-Verlag.

[35] Paganini, F, Tang, A., Ferragut, A. and Andrew, L.L.H. Network stability under alpha fair bandwidth allocation with general file size distribution. *IEEE Transactions on Automatic Control*, 57, No. 3 (2012) 579591.

[36] Perthame, B. "*Transport equations in biology*," (2007), Frontiers in Mathematics Series, Birkhäuser.

[37] Puha, A.L. and Ward, A. Scheduling an overloaded multiclass many-Server queue with impatient customers, in *Tutorials in Operations Research: Operations Research & Management Science in the Age of Analytics*, (2019) 189217.

[38] Puha, A.L. and Williams, R.J. Asymptotic behavior of a critical fluid model for a processor sharing queue via relative entropy, *Stochastic Systems*, 6 (2016), 251-300.

[39] Smith, W.L. Asymptotic renewal theorems. *Proc. Roy. Soc. Edinb. A* 64 (1954) 9-48.

[40] Whitt, W. Fluid models for multi-server queues with abandonments. *Operations Research*, 54, No. 1 (2006) 37-54.

[41] Yu, Y. Concave renewal functions do not imply DFR interrenewal times, *J. Appl. Probab.*, 48, No. 2 (2011), 583-588.

[42] Zeltyn, S. and Mandelbaum, A. Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queueing Syst.* 51 (2005) 361402.

Viterbi Faculty of Electrical and Computer Engineering, Technion, Haifa, Israel

*Email address*: rami@technion.ac.il

Department of Mathematics and Statistics,, University of Maryland, Baltimore County

*Email address*: wkang@umbc.edu

Department of Industrial Engineering and Management, Technion, Haifa, Israel

*Email address*: hkaspi@technion.ac.il

Division of Applied Mathematics, Brown University, Providence, RI 02118, USA

*Email address*: kavita_ramanan@brown.edu