SIAM J. MATH. ANAL.
Vol. 55, No. 6, pp. 7189–7239

LONG-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 λ, is considered. These equations describe the hydrodynamic or fluid limit of many-server queues with reneging (with traffic intensity λ), 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 nonincreasing 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 invariant state is the unique invariant distribution of the fluid equations. In particular, this implies that the stationary distributions of scaled N-server systems converge to the unique invariant state of the corresponding fluid equations. Moreover, when the mean arrival rate is not equal to the mean service rate, that is, 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, as well as under the two regimes $\lambda < 1$ and $\lambda \geq 1$. When the hazard rate function is nonincreasing, 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.

Key words. 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

MSC codes. 60F17, 60K25, 90B22, 60H99, 35D99

DOI. 10.1137/21M1433125

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. Such queues arise in a range of applications, including as models of computer networks, telephone call

[*] Received by the editors July 12, 2021; accepted for publication (in revised form) July 12, 2023; published electronically November 7, 2023.
https://doi.org/10.1137/21M1433125
Funding: The work of the first author was supported by Israel Science Foundation grant 1035/20. The work of the fourth author was supported by U.S. Army Research Office grant W911NF2010133 and the Office of Naval Research under Vannevar Bush Faculty Fellowship grant N0014-21-1-2887.

[†] Viterbi Faculty of Electrical and Computer Engineering, Technion, Haifa, Israel (rami@technion.ac.il).
[‡] Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250 USA (wkang@umbc.edu).
[§] Department of Industrial Engineering and Management, Technion, Haifa, Israel (iehaya@technion.ac.il).
[¶] Division of Applied Mathematics, Brown University, Providence, RI 02118 USA (Kavita_Ramanan@brown.edu).

centers or (more general) customer contact centers [18, 31, 42], and enzymatic processing networks in biology, where reneging seeks to model the phenomenon of dilution (see, e.g., [32]). 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 renege from (equivalently, abandon) the queue at the moment when the amount of time they have been waiting in queue equals their patience time, unless they have already entered service by that 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 reneging or abandonment distribution is general, explicit formulas for the scaled steady-state distributions were obtained in [11], and their asymptotics as N, the number of servers, goes to infinity, were studied in [45]. However, the case of general service distributions, which is relevant for many applications, is more challenging. 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 [26] 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 ν^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. Since the number of servers is N, it follows that the total mass $\nu^N[0,\infty)$ of ν^N, which represents the number of busy servers, is less than or equal to N. For analytical purposes, it turns out that the queue measure itself is more conveniently represented in terms of the pair (η^N, X^N), where η^N is a potential queue measure, which keeps track of the times elapsed since entry into the system, not only of jobs currently in the queue, but of all jobs that have entered the system and for which this elapsed time is strictly less than their respective patience time (regardless of whether or not they entered service or departed the system by that time), and X^N is the total number of jobs in the system. The work [26] considered a general class of service and patience distributions, namely those whose cumulative distribution functions, denoted G^s and G^r, respectively, have finite means, have densities, denoted by g^s and g^r, respectively, and whose associated hazard rate functions $h^s = g^s/(1-G^s)$ and $h^r = g^r/(1-G^r)$ satisfy mild additional regularity conditions (see Assumption 2.1). For this class, it was shown in [26] that when the traffic intensity, that is, the ratio of arrival rate to number of servers, converges to a limit $\lambda \geq 0$ as the number of servers tends to infinity, the rescaled state descriptor $(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N) := N^{-1}(X^N, \nu^N, \eta^N)$ converges to a deterministic limit (X, ν, η), where for each $t \geq 0$, $X(t)$ is a nonnegative number representing the limiting scaled number of jobs in queue, ν_t is a subprobability measure on $[0,\infty)$ (i.e., with mass no greater than 1) representing the scaled limit distribution of ages of jobs in service, and η_t is a finite nonnegative Borel measure on $[0,\infty)$

representing the scaled limit distribution of times since entry into the system of jobs whose times in the system are strictly less than their patience times. Moreover, this deterministic limit (X, ν, η) was characterized as the unique solution to a system of coupled equations that we refer to as the *fluid equations* (see Definition 2.3) which, in particular, characterize ν and η as the unique weak solution to a nonlinearly coupled system of deterministic measure-valued transport equations, subject to discontinuous boundary conditions that are modulated by the state of X.

In this work we study the large-time behavior of the solution (X, ν, η) to the measure-valued fluid equations obtained in [26] under the additional assumption that the fluid equations admit a unique invariant state (sometimes also referred to as a fixed point). Although a fully rigorous definition of these equations is somewhat involved and hence deferred to section 2.2 (specifically, see Definition 2.3), to relate our model to the literature on age-structured population dynamics, as well as to illustrate some of the challenges, we briefly describe a simplified version of the dynamics of the measure-valued component ν under additional smoothness assumptions. Suppose that for each $t \geq 0$, the server age measure ν_t also has a continuous density, denoted by $f(\cdot, t)$. A purely formal derivation using the fluid equations in Definition 2.3 shows that the server age measure density f satisfies the following transport partial differential equation (PDE) (see section 4.2.2):

$$\partial_t f(x,t) + \partial_x f(x,t) + h^s(x) f(x,t) = 0, \quad x > 0, t > 0, \tag{1.1}$$

subject to the boundary condition

$$f(0,t) = \begin{cases} \lambda & \text{if } \displaystyle\int_0^\infty f(x,t)dx < 1, \\ \displaystyle\int_0^\infty h^s(x) f(x,t)dx & \text{if } \displaystyle\int_0^\infty f(x,t)dx = 1, \end{cases} \tag{1.2}$$

and initial condition

$$f(x,0) = f_0(x), \quad x > 0, \tag{1.3}$$

for some given function $f_0 : \mathbb{R}_+ \to \mathbb{R}_+$ with $\int_0^\infty f_0(x) \leq 1$. Recall that $\int_0^\infty f(x,t)dx$ is the limiting fraction of busy servers, and note that the discontinuity in the boundary condition (1.2) reflects the difference in the rates of entry of jobs into service when all servers are busy and when there is a positive fraction of idle servers. As a word of caution we mention that uniqueness is not in general guaranteed for (1.1)–(1.3) when they are considered alone. However, as already remarked above, in the full formulation, ν is coupled to the other components, and uniqueness as well as the large-time behavior rely on this coupling. Nevertheless, a discussion of just this PDE will still provide insight into some of the challenges that arise in the study of the large-time behavior of the full system of fluid equations.

The PDE (1.1) is reminiscent of age-structured equations (sometimes also referred to as renewal equations in the PDE literature) that model population dynamics in biology. For example, if (1.1) is accompanied by the boundary condition $f(0,t) = \int_0^\infty b(x) f(x,t)dx$ in place of (1.2), it provides a model for the evolution of a population where birth and death rates are age-dependent. In this model, individuals of age x give birth at rate $b(x)$ and die at rate $h^s(x)$, and f describes the population density with respect to age. This is a fully linear equation about which much is known. In particular, the decay or growth rates as well as the large-time profile are given in terms of an eigenvalue problem (e.g., see [39, Chapter 2]). There is a large body of literature on various generalizations of the PDE (1.1), for example, with additional linear or nonlinear terms on the right-hand side, and their use in biological modeling. We only mention a small sample of works that are somewhat related to ours, and refer the interested reader to monographs for further information and references [16, 39, 40, 41]. For example, the work [33] studies the long-time behavior of several linear structured population models using a combination of eigenvalue problems and generalized relative entropy techniques, whereas the work [21] is concerned with measure-valued solutions and shows that a variant of the Monge–Kantorovich transport distance is nonexpanding for these equations. The nonlinear theory is far more involved because different nonlinearities, and sometimes even different data, lead to different behaviors, including chaotic, periodic, and convergence to a steady state at the large-time limit. The work [13] considers a class of nonlinear selection-mutation and structured population models and is also concerned with measure-valued solutions, motivated by the fact that some of the invariant distributions do not have densities with respect to Lebesgue measure. However, the latter work focuses on well-posedness rather than large-time behavior. The paper [34] studies several variants of linear and nonlinear population models that in addition to being age-structured allow for the rate of maturation (or aging) to vary across the population. The rate of exponential growth and limiting profile, and conditions for existence of nonlinearly stable states as well as oscillatory behavior are provided. All these works consider linear boundary conditions, unlike that in (1.2). The work on age-structured models where birth processes are described via nonlinear boundary conditions is much less extensive, but we mention [36], which studies a model with spatial diffusion and proves well-posedness, and [19], which addresses a specific type of nonlinearity aimed at modeling competition between young and old parts of the population and establishes large-time convergence results in the presence of singularly perturbed dynamics. The type of discontinuous nonlinearity we are concerned with is not covered by any of these papers.

In a broad sense, (1.1)–(1.3) can be seen as belonging to the same family of evolution equations describing age-dependent population dynamics. However, there are significant differences. First, to further illustrate the difference in the boundary conditions, consider a simplified, linear version of (1.2) where $f(0,t) = \int_0^\infty h^s(x) f(x,t)dx$ holds for all $t \geq 0$. Then one recovers, for example, the age-structured model (3) in [33] with $\nu \equiv 0$ and $b = d = h^s$. In this case it follows from [34, 33] that $f(x,t)$ converges to the eigenfunction $f_*(x) := 1 - G^s(x)$ of the stationary version of the PDE (1.1) (corresponding to the eigenvalue 0) when h^s is strictly positive, bounded, and integrable on $(0,\infty)$. However, this result is not applicable in the setting of the present paper where h^s represents a hazard rate function, because the hazard rate function of any distribution is never integrable on $(0,\infty)$; note that this also implies that the dual problem of the age-structured model in equation (8) of [33] is not well defined. More significantly, in the current setting a key additional challenge, especially in the critical and supercritical regimes where $\lambda \geq 1$, is to deal with the discontinuous boundary condition (1.2), which requires controlling the oscillations of $\int_0^\infty f(x,t)dx$ around 1. Furthermore, a rigorous proof cannot in any case rely on just the analysis of the PDE because for general initial conditions ν_0, the measures ν_t, $t \geq 0$, need not have densities and even when they do, their densitites will have discontinuities in both variables (as is apparent from the analysis in section 4.2.3). Finally, as already mentioned, in addition to the measure-valued analogue of (1.1)–(1.3), we need to consider the dynamics of the additional components η and X of the full state of the system. Thus, we develop alternative methods to analyze these fluid equations.

In this paper, we study the subcritical, critical, and supercritical regimes, characterized by the regions where $\lambda < 1$, $\lambda = 1$, and $\lambda > 1$, respectively, additionally assuming in the critical and supercritical regimes that the hazard rate function of the service time distribution is either nonincreasing or bounded away from zero and infinity. Our main results are summarized in Theorem 3.2. Specifically, when $\lambda \neq 1$, we show that 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 nonincreasing, we show that the total mass of ν (which represents the mass of busy servers in the fluid system) converges to 1. In all the cases above, we show that the Dirac delta mass at the (unique) invariant state is the unique *invariant distribution* of the fluid equations (i.e., the unique probability distribution that is invariant under the flow defined by the fluid equations; see Definition 2.10). This crucially implies that the stationary distributions of the rescaled 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 3.4, it is the uniqueness of the invariant distribution for the fluid equations, rather than just uniqueness of the invariant state, that is relevant for the convergence of the rescaled stationary distributions of the N-server dynamics. In the absence of reneging, such large-time convergence results were established for a single-class system in Proposition 6.1 of [28] for the subcritical regime and in Theorem 3.9 of [28] 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.

Finally, we also analyze the large-time behavior of fluid equations for a multiclass model under a nonpreemptive priority policy, which was formulated in [10] and used therein to establish asymptotic optimality of the policy when the reneging distribution is exponential (see Definition 5.1). In the case that the 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 invariant distribution and analogous large-time convergence results in the supercritical regime.

While [28] is one of the earliest works to establish large-time limits of measure-valued fluid equations in the queueing context, an early work that used a Lyapunov functional approach to establish large-time behavior of measure-valued fluid equations is [38], and the use of relative entropy as a Lyapunov functional seems to have first appeared in [43], with subsequent generalizations considered in [22, 35]. Another work that addresses large-time behavior for measure-valued fluid equations, although using a different (non-Lyapunov) method, is [29]. However, with the exception of [28], all of these works focus on the dynamics of residual times (specifically, for jobs in processor sharing, bandwidth sharing, and GI/G/N+G models), which have a different structure from the equations studied in this paper.

1.2. Discussion of the proof methodology. For the single-class setting, 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 two different types of assumptions on the service distribution (see Asssumption 3.1). When the hazard rate function of the service distribution is nonincreasing, we use a reformulation in terms of renewal equations in the probabilistic sense, 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 aforementioned work [29], 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 nonincreasing hazard rate functions). However, as already mentioned above, the fluid equations of [29] are based on a different measure-valued state representation, involving residual service times of customers rather than ages, and furthermore, convergence is established in [29] only for the queue process, not the measure-valued process. Thus the results of [29] do not directly apply. Moreover, we also need to establish additional estimates to prove convergence of the measure-valued process ν in the supercritical case, which is not addressed in [29].

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 [29]. They are based on the analysis of weak solutions to PDEs and entail showing that an extended relative entropy functional (that takes as arguments subprobability 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. Thus, these methods are closer in spirit to the methods developed to study age-structured population models alluded to above, but with important differences to address the additional complications. To further elaborate on the additional subtleties that arise in this context we refer back to the formal PDE for the density f given in (1.1)–(1.3), and the function $f_* = 1 - G^s$, and assuming (again without justification) that $\lim_{x\to\infty} f(x,t)\log(f(x,t)/f_*(x)) = 0$ for all $t \geq 0$, consider the extended relative entropy functional

$$\hat{\Delta}(t) := \int_0^\infty f(x,t)\log\frac{f(x,t)}{f_*(x)}dx, \qquad t \geq 0.$$

Note that unlike usual relative entropy, $\hat{\Delta}(t)$ may be negative since $x \mapsto f(x,t)$ need not be a probability density. Then using (1.1)–(1.3), combined with several other estimates and manipulations (see section 4.2.2 for full details), it is possible to show that

$$\text{(1.4)} \qquad \frac{d}{dt}\hat{\Delta}(t) \leq \begin{cases} -\varepsilon^s\hat{\Delta}^+(t) + \lambda\log\lambda - z(t)\log z(t) & \text{if } \int_0^\infty f(x,t)dx < 1, \\ -\varepsilon^s\hat{\Delta}(t) & \text{if } \int_0^\infty f(x,t)dx = 1, \end{cases}$$

where $a^+ := \max(a,0)$ and $z(t) := \int_0^t h^s(x)f(x,t)dx$ and $\varepsilon^s := \operatorname{ess\,inf}_{x\geq 0} h^s(x)$. Given $\varepsilon^s > 0$, this suggests that $\hat{\Delta}(t)$ may serve as a sort of Lyapunov function for the dynamics if there exists a finite time $T < \infty$ such that $\int_0^\infty f(x,t)dx = 1$ for all $t \geq T$. However, translating this intuition into a proof is not at all straightforward. For one, showing that such a time T has been reached cannot be based on (1.4) alone, because the dynamics could oscillate between times when $\int_0^\infty f(t,x)dx = 1$, whence $\hat{\Delta}(t) \geq 0$, and time intervals when $\int_0^\infty f(t,x)dx < 1$, whence $\hat{\Delta}(t)$ could be negative. As a consequence, it is not possible to deduce from $\hat{\Delta}(t) \leq 0$ that $t \geq T$. Using the coupling with dynamics of the other state components is necessary. Second, the calculation (1.4) was carried out under the unjustified assumption that $\lim_{x\to\infty} f(x,t)(\log f(x,t)/f_*(x)) = 0$ for all $t \geq 0$. Third, as already mentioned above, due to lack of sufficient smoothness one cannot analyze just the PDE but one must consider the full measure-valued equation. Nevertheless, in the supercritical regime $\lambda > 1$, under the assumption that both $\varepsilon^s > 0$ and $c^s := \operatorname{ess\,sup}_{x>0} h^s(x) < \infty$, stated as Assumption 3.1(1), below, we first show convergence of ν_t to the measure $\nu_*(dx) = f_*(x)dx$ as $t \to \infty$ via a more involved rigorous argument. The latter first derives a version of (1.4) for the measure-valued counterpart ν_t, which contains some extra error terms, then uses that to deduce the existence of a finite time T such that the total mass of ν_t is 1 for $t \geq T$, and then finally deduces the convergence of the full state (X_t, ν_t, η_t) to the unique invariant state; see section 4.2.3 for full details.

1.3. Ramifications of our results for stationary distributions of N-server queues. The results of this paper also give 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 Theorems 3.2 and 7.1 of [27] that the measure-valued state dynamics (X^N, η^N, ν^N) for each N-server system describe an ergodic Feller process with a unique stationary distribution, whereas Theorem 3.3 of [27] shows that the sequence of stationary distributions $\{\bar{\pi}^N\}_{N\in\mathbb{N}}$ of the normalized states $(\bar{X}^N, \bar{\eta}^N, \bar{\nu}^N) = N^{-1}(X^N, \eta^N, \nu^N)$, $N \in \mathbb{N}$, is tight. Moreover, the latter theorem also states that any subsequential limit of $\{\bar{\pi}^N\}_{N\in\mathbb{N}}$ must coincide with the Dirac delta mass at 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 [27]. One only knows that any subsequential limit of the scaled stationary distributions of N-server queues with reneging $\{\bar{\pi}^N\}_{N\in\mathbb{N}}$ is an invariant distribution of the fluid equations (see Definition 2.10), and a priori one does not know that this distribution is concentrated on the unique invariant state of the fluid equations. However, as shown in Proposition 4.16 of the present paper, when from any initial condition there is convergence, as time tends to infinity, of the fluid equations' solution to the unique invariant state (or, when $\lambda = 1$, just convergence of η_t and the fraction of busy servers to one as $t \to \infty$), it follows that the set of invariant distributions has just one element, equal to the Dirac delta measure at the unique (deterministic) invariant state of the fluid equations, thus closing the gap in the proof of Theorem 3.3 of [27] under additional assumptions.

Our work in the multiclass setting also closes an exactly analogous gap in the proof of Theorem 4.4 of [10] under additional assumptions. 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 3.4 for further elaboration of this point). In the case of [10], the gap also affects the validity of Theorem 5.1 therein, regarding the asymptotic optimality of an index policy, referred to as the $c\mu/\theta$ rule, which was introduced in [8]. The results obtained in this paper validate the asymptotic optimality result in Theorem 5.1 of [10] 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 [8] and [9], which deal with the case of exponential service time distributions; recent developments on this policy under various additional settings can also be found in [30].) Finally, we note that limits of stationary distributions of many-server systems in the (so-called Halfin–Whitt) diffusive regime have been considered in [23, 3, 4] in the absence of reneging and in [25, 17, 24] in the presence of reneging.

1.4. Open problems. This work leads to several interesting open problems. One problem worthy of future investigation would be to determine precisely the full class of service distributions for which such large-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 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 PDEs [5], and uniqueness of the invariant state 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 one to treat asymptotic optimality of the aforementioned index rule in broader settings. It would also be interesting to see if the methodology developed here can be used to analyze certain classes of age-structured or size-structured population models. It is worth mentioning that this model has a loose connection to the model treated in [34], which allows for different maturation rates for different parts of the population (namely, customer classes).

1.5. Organization of the rest of the paper. In section 2.2 we introduce the fluid equations in the single-class setting, and in section 2.3 define invariant states of the fluid equations. 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 1.6, we introduce common notation that is used throughout the paper.

1.6. 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, and $\mathbb{R}_+$ is the set of nonnegative real numbers. For $a, b \in \mathbb{R}$, $a \vee b$ denotes the maximum of a and b, $a \wedge b$ the minimum of a and b and the shorthand 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 $\|f\|_\infty := \sup_{x\in[0,\infty)} |f(x)|$, which could possibly take the value ∞. In addition, the support of a function f is denoted by $\mathrm{supp}(f)$. Given a nondecreasing function f on $[0,\infty)$, f^{-1} denotes the inverse function of f, defined precisely as

$$f^{-1}(y) = \inf\{x \geq 0 : f(x) \geq y\}. \tag{1.5}$$

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 ψ to denote generic functions on $[0,H)$ and φ to denote generic functions on $[0,H)\times\mathbb{R}_+$. By some abuse of notation, given ψ 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 $\mathcal{M}(E)$ to denote the space of Radon measures on a metric space E, endowed with the Borel σ-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 δ_x will be used to denote the measure with unit mass at the point x and, with some abuse of notation, we will use $\mathbf{0}$ 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 shorthand 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 shorthand 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 σ-algebra.

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

$$F^\mu(x):=\mu[0,x],\quad x\in[0,H), \tag{1.6}$$

and we define $(F^\mu)^{-1}$ to be its left-continuous inverse as defined in (1.5), that is,

$$(F^\mu)^{-1}(y)=\inf\{x>0:F^\mu(x)\geq y\}. \tag{1.7}$$

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 as $t\to\infty$:

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

We will also on occasion use the total variation distance on $\mathcal{M}_F[0,H)$, denoted by $d_{\text{TV}}(\mu,\nu):=2\sup_{A\in\mathcal{F}}|\mu(A)-\nu(A)|$, where $\mathcal{F}$ is the Borel σ-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 let $\mathcal{I}_{\mathbb{R}_+}[0,\infty)$ denote the subset of nondecreasing 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. The N-server model, fluid equations, and invariant distributions.

2.1. Description of the N-system. We start by briefly describing the N-server model dynamics, as introduced in [27]. 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. 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,

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

and we normalize units so that the mean service time is 1, namely,

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

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. The mild, but somewhat technical, condition in part (2) of the assumption is required in order to use various results from [26, 28], where it has been assumed. Note that, strictly speaking, g^s and g^r (and thus h^s and h^r) are determined only almost everywhere (a.e.). The convention implicitly adopted in the above statement is that h^s (respectively, h^r) is a.e. equal to a function from $\mathbb{R}_+$ to itself that is bounded or lsc.

For each $N \in \mathbb{N}$, we consider a system in which jobs with i.i.d. patience times $(r_j)_{j\in\mathbb{Z}}$ and i.i.d. service times $(v_j)_{j\in\mathbb{Z}}$, both mutually independent of each other, arrive to a system of N servers and are served in the order of arrival. Jobs that arrive when there is an idle server present immediately enter service (at a server chosen at random from among the idle servers), while jobs that arrive when all servers are busy wait in queue in the order of their arrival. Jobs renege from the queue at the moment when their time in queue equals their patience time, and jobs that reach the head of the queue and do not renege before a server becomes available enter service at the moment when this server becomes available. To describe the dynamics, let $E^N(t)$ denote the cumulative number of jobs that arrived into the N-system in the interval $[0,t]$ and let $\mathcal{I}^N(t)$ represent the set of indices of jobs that entered the N-system by time t (which includes the positive indices $j = 1,\ldots,E^N(t)$ of jobs that arrived after time 0 as well as certain nonpositive indices representing jobs present in the system at time 0). For each job $j \in \mathcal{I}^N(t)$, let $w_j^N(t)$ represent its potential waiting time, or the amount of time prior to t since it entered the system. Then, for $t \geq 0$, let η_t^N represent the potential queue measure at time t, which has a Dirac delta mass at the potential waiting times of jobs that entered the system but did not renege from the system by time t:

$$\eta_t^N = \sum_{j\in\mathcal{I}^N(t)} \delta_{w_j^N(t)} 1_{\{w_j^N(t) < r_j\}},$$

where the indicator functions ensure that only jobs whose patience is strictly less than their waiting time at time t are retained in the sum. Note that each η_t^N is a random element taking values in the space of nonnegative measures on $[0,H^r)$. Similarly, for each job $j\in\mathcal{I}^N(t)$, let $a_j^N(t)$ denote its age at time t, which is equal to zero if it has not entered service by time t, or equal to the amount of time elapsed since the job j entered service if it is still in service at time t, or equal to its service time v_j if it has departed from the system by time t. Then the server age measure ν_t^N, which has a Dirac delta mass at the age of each job that is in service at time t, can be written as

$$\nu_t^N=\sum_{j\in\mathcal{I}^N(t)}\delta_{a_j^N(t)}1_{\{\frac{da_j^N}{dt}(t+)>0\}},$$

where the indicator functions in the sum serve only to select indices of those jobs j that are in service at time t, which can be characterized by the condition $\frac{da_j^N}{dt}(t+)>0$. Note that each ν_t^N is a random element taking values in the space of nonnegative measures on $[0,H^r)$ whose total mass $\langle\mathbf{1},\nu_t^N\rangle$ is no greater than N. Next, let $X^N(t)$ be the nonnegative real-valued random variable that represents the number of jobs in the system at time t. Then, since the N-system is assumed to be nonidling in the sense that one can never have an idle server when there is a job in queue, this implies that whenever the total number of jobs in system is greater than N, there are no idle servers, that is,

$$N-\langle\mathbf{1},\nu_t^N\rangle=[N-X^N(t)]^+. \tag{2.5}$$

Moreover, clearly $Q^N(t)=X^N(t)-\langle\mathbf{1},\nu_t^N\rangle$ represents the number of jobs waiting in queue at time t.

If E^N is a Poisson process, then it is not hard to see that (X^N,ν^N,η^N) is a Markov process taking values in $\mathbb{R}\times\mathcal{M}_F[0,H^s)\times\mathcal{M}_F[0,H^r)$ that describes the state of the system. The full characterization of the dynamics of the N-system given in [26] also involved the auxiliary processes $K^N(t)$, $D^N(t)$, $R^N(t)$, and $S^N(t)$ that denote the total number of jobs that, respectively, entered service, departed from the system (on completing service), and reneged from the queue, and whose time since entry into service exceeded their patience time during the interval $[0,t]$. $S^N(t)$ is referred to as the potential cumulative reneging process since it also counts jobs that may have entered service or departed the system at the time their patience time became equal to the time since entry into the system. As shown in [26, Theorem 2.1], η^N and ν^N satisfy certain measure-valued transport equations driven by the processes E^N and K^N, respectively. For each $N\in\mathbb{N}$ and $H^N=E^N,D^N,K^N,R^N,S^N,Q^N,X^N,\eta^N,\nu^N$, define $\bar{H}^N:=H^N/N$. Assume that for each $N\in\mathbb{N}$, E^N is a Poisson process with rate λ^N, and $\lambda^N/N\to\lambda\in(0,\infty)$ as $N\to\infty$. It was shown in [26] that as $N\to\infty$, the random process $(\bar{X}^N,\bar{\nu}^N,\bar{\eta}^N)$ converges weakly to a deterministic limit (X,ν,η) that is the unique solution to a coupled system of equations referred to as the fluid equations with arrival rate λ, which are defined in the next section.

2.2. Fluid equations.

We now introduce the fluid equation that describes the limit of the scaled N-server state process $(\bar{X}^N,\bar{\nu}^N,\bar{\eta}^N)$ introduced in the last section. When referring to these limit objects we will not use the term *jobs* but rather *mass,* which is more suitable for their continuous counterpart. However, with some abuse of terminology, the terms *fraction of servers that are busy, fraction of servers that are idle,* etc., will be used when referring to the fluid counterparts of these N-server model processes.

The fluid equations are concerned with a triplet (X, ν, η), where $X(t)$ represents the total mass in the system at time t, including mass in the queue and mass in service, ν_t is the fluid age measure, which is a subprobability measure on $[0, H^s)$ that assigns to any interval $[a, b) \subset [0, \infty)$ the (limiting) fraction of servers for whom the mass currently in service has been in service for a number of time units lying in $[a, b)$, and η_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 that has arrived by time t and whose potential waiting time at time t lies in $[a, b)$ (irrespective of whether or not they have entered service or departed the system by time t), as long as it has not reneged from the queue by time t. Note that the total fraction of idle 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 $1 - \langle \mathbf{1}, \nu_t \rangle = [1 - X(t)]^+$, which is the scaling limit analogue of (2.5).

The input data for the fluid equations includes the limiting arrival rate λ 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

$$\mathfrak{S} := \left\{ \begin{matrix} (\tilde{x}, \tilde{\nu}, \tilde{\eta}) \in \mathbb{R}_+ \times \mathcal{M}_F[0, H^s) \times \mathcal{M}_F[0, H^r) : \\ 1 - \langle \mathbf{1}, \tilde{\nu} \rangle = [1 - \tilde{x}]^+ \end{matrix} \right\}. \tag{2.6}$$

We now give a precise formulation of the fluid equations introduced in [27] with $E(t) = E^\lambda(t) := \lambda t$ for $t \geq 0$ therein, where λt represents the total mass to arrive by time t, and subsequently provide an intuitive explanation of the form it takes. These equations will also involve the auxiliary processes mentioned earlier, namely the queue mass process $Q(\cdot)$, and the nondecreasing processes $D(\cdot), K(\cdot)$, $S(\cdot)$, and $R(\cdot)$. Here, $Q(t)$ represents the total mass in queue (awaiting service) at time t, and $D(t), K(t)$, $S(t)$, and $R(t)$ represent, respectively, the cumulative mass of departures from the system on completion of service, mass of entries into service, mass of potential abandonments from the system, and mass of actual abandonments from the system in the interval $[0, t]$.

Definition 2.3 (fluid equations). *Given $\lambda \geq 0$ and hazard rate functions h^r and h^s, the càdlàg function (X, ν, η) 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*

$$S(t) := \int_0^t \langle h^r, \eta_u \rangle \, du < \infty, \qquad D(t) := \int_0^t \langle h^s, \nu_u \rangle \, du < \infty, \tag{2.7}$$

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

$$\begin{aligned} \langle \varphi(\cdot, t), \nu_t \rangle = \langle \varphi(\cdot, 0), \nu_0 \rangle &+ \int_0^t \langle \varphi_u(\cdot, u) + \varphi_x(\cdot, u), \nu_u \rangle \, du \\ &- \int_0^t \langle h^s(\cdot) \varphi(\cdot, u), \nu_u \rangle \, du + \int_0^t \varphi(0, u) \, dK(u), \end{aligned} \tag{2.8}$$

where

$$K(t) = \langle \mathbf{1}, \nu_t \rangle - \langle \mathbf{1}, \nu_0 \rangle + D(t); \tag{2.9}$$

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

$$\langle \varphi(\cdot,t),\eta_t\rangle = \langle \varphi(\cdot,0),\eta_0\rangle + \int_0^t \langle \varphi_u(\cdot,u)+\varphi_x(\cdot,u),\eta_u\rangle\, du \\ - \int_0^t \langle h^r(\cdot)\varphi(\cdot,u),\eta_u\rangle\, du + \lambda\int_0^t \varphi(0,u)\, du; \tag{2.10}$$

with the nonidling constraint

$$1-\langle \mathbf{1},\nu_t\rangle = [1-X(t)]^+, \tag{2.11}$$

where

$$X(t) = X(0)+\lambda t - D(t) - R(t), \tag{2.12}$$

with

$$R(t) = \int_0^t \left(\int_0^{Q(u)} h^r((F^{\eta_u})^{-1}(y))dy \right) du, \tag{2.13}$$

where recall $F^{\eta_t}(x) := \eta_t[0,x]$*, and* $(F^{\eta_t})^{-1}$ *denotes the left-continuous inverse defined in* (1.7)*, and*

$$Q(t) = X(t) - \langle \mathbf{1},\nu_t\rangle, \tag{2.14}$$

with Q *also satisfying the inequality constraint*

$$Q(t) \le \langle \mathbf{1},\eta_t\rangle. \tag{2.15}$$

Remark 2.4. Note that if (X,ν,η) solves the fluid equations with arrival rate λ 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 η_t in (2.19) below, from which it is clear that if η_0 does not charge points, then neither does η_t).

Next, note from (2.14) and (2.11) that for each $t\in[0,\infty)$,

$$Q(t) = [X(t)-1]^+. \tag{2.16}$$

For future use, we also observe that (2.9), (2.14), and (2.12), when combined, show that for every $t\in[0,\infty)$,

$$Q(0)+\lambda t = Q(t)+K(t)+R(t), \tag{2.17}$$

which is simply a mass conservation equation upon recalling the interpretation of λt, $Q(t), K(t)$, and $R(t)$ above Definition 2.3. In addition, we will find it convenient to define

$$B(t) := \langle \mathbf{1},\nu_t\rangle, \quad t\ge 0, \tag{2.18}$$

which represents the limiting fraction of busy servers.

Remark 2.5. Given a solution (X,ν,η), we will refer to (B,D,K,Q,R,S) as auxiliary processes. In addition to the term fluid equations we will also use the term fluid system to refer to a tuple (X, ν, η) along with the auxiliary processes (B, D, K, Q, R, S) satisfying the fluid equations.

We now provide an informal, intuitive explanation for the form of the fluid equations, which also aligns with the description of the N-system given in section 2.1. First, $\nu_u(dx)$ represents the fraction of servers that are processing mass whose age lies in the range $[x, x+dx)$ at time u, and $h^s(x)$ represents the conditional mean rate at which mass with age in $[x, x+dx)$ completes service given that its age is at least x. Hence, in (2.7), $\langle h^s, \nu_u\rangle$ represents the departure rate of mass due to services at time u, and $D(t)$, its integral over $[0,t]$, is the total departure due to service completion in the interval $[0,t]$. By an exactly analogous reasoning, the other quantity $S(t) = \int_0^t \langle h^r, \eta_u\rangle du$ in (2.7) represents the cumulative potential reneging from the system in the interval $[0,t]$. However, the actual reneging rate is restricted to abandonments of the mass in the queue. Since entry into service takes place in the order of arrival, the age of the oldest (equivalently, head-of-the-line) mass in the queue is $\bar{a}_u := (F^{\eta_u})^{-1}(Q(u))$, so that $\eta_u[0, \bar{a}_u] = Q(u)$. Here, recall that F^{η_u} represents the cumulative distribution function of the measure η_u (which need not be a probability measure). Thus, the actual reneging rate at any time u only counts the mass reneging from the potential queue measure η_u whose age lies in the restricted interval $[0, \bar{a}_u]$, rather than the entire interval $[0, \infty)$. A standard change of variables then yields the expression in (2.13). Next, recalling the interpretations of the quantities K, R, and Q stated prior to Definition 2.3, note that (2.9), (2.14), and (2.12) are simply mass conservation equations, and (2.11) represents a nonidling condition that ensures that no server can idle when there is work in the queue. Moreover, the inequality (2.15) expresses the constraint that at any time t, the mass in the queue is bounded by the total mass of the 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, (2.8) and (2.10) govern the evolution of the fluid age measure ν and potential queue measure η, respectively. In particular, the second term on the right-hand side of (2.8) 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 the system, which are driven by the function K, the cumulative entry into service. Equation (2.10) is exactly analogous, but with h^r and the cumulative arrivals E^λ 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 most of which has been proved in [26, 28]. Recall the definition of the space $\mathfrak{S}$ given in (2.6).

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* λ *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* λ *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* λ *and* $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$, *the following properties hold:*

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

$$\langle \psi, \eta_t \rangle = \int_{[0,H^r)} \psi(x+t) \frac{\bar{G}^r(x+t)}{\bar{G}^r(x)} \eta_0(dx) + \int_0^t \psi(u) \bar{G}^r(u) \lambda du; \tag{2.19}$$

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

$$\langle \psi, \nu_t \rangle = \int_{[0,H^s)} \psi(x+t) \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + \int_0^t \psi(t-u) \bar{G}^s(t-u) dK(u), \tag{2.20}$$

with K equal to the auxiliary process defined in (2.9) *of the fluid equations;*

(iii) *if Q and B are the associated auxiliary processes defined in* (2.14) *and* (2.18), *respectively, then K is an absolutely continuous function and the derivative K' of K satisfies for a.e. $t \geq 0$,*

$$K'(t) = k(t) := \begin{cases} \lambda & \text{if } B(t) < 1, \\ \langle h^s, \nu_t \rangle & \text{if } B(t) = 1. \end{cases} \tag{2.21}$$

Proof. Uniqueness of the solution to the fluid equations follows from [26, Theorem 3.5] 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 λ 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 [26, Theorem 3.6], once we justify that the conditions of that theorem are satisfied in the present setting. First, it is not hard to see that for the fixed $\lambda \geq 0$ and $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$, 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 [26, Assumption 3.1] is satisfied. Second, note that since η_0 is a continuous measure, and E^λ is continuous, [26, Assumption 3.2] is also satisfied. Finally, [26, Assumption 3.3] is a direct consequence of Assumption 2.1 of this paper, and thus the application of [26, Theorem 3.6] is justified.

We now turn to estabishing the properties of any solution (X, ν, η) to the fluid equations. First, note that the forms of both (2.10) and (2.8) are analogous to that of (4.2) in [28], and therefore the integrability conditions in (2.7) imply that (4.1) of [28] holds. Thus, (2.19) and (2.20) for $\psi \in \mathcal{C}_c[0, H^r)$ and $\psi \in \mathcal{C}_c[0, H^s)$, respectively, follow from [28, Theorem 4.1]. 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 ψ. In particular, these equations also hold with $\psi = h^r$ in (2.19) and $\psi = h^s$ in (2.20). 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 formula (3.12) of [28] and Corollary 3.7 of [26]. We supply the details for completeness. First, note that D, S, and R are absolutely continuous by definition. Thus, (2.12) shows that X is absolutely continuous. Thus, since $B(t) = \langle \mathbf{1}, \nu_t \rangle = \min(X(t), 1)$ for $t \geq 0$, B is also absolutely continuous. In turn, by (2.9) and (2.14), this implies that K is absolutely continuous. Further, (2.9), (2.17), (2.7), and (2.18) show that for a.e. $t > 0$,

$$K'(t) = \lambda - Q'(t) - \int_0^{Q(t)} h^r((F^{\eta_t})^{-1}(y)) dy, \qquad \text{and} \qquad K'(t) = B'(t) + \langle h^s, \nu_t \rangle. \tag{2.22}$$

We now recall the fact that given $c \in \mathbb{R}$ and an absolutely continuous function $f : \mathbb{R} \to \mathbb{R}$, denoting by f' its derivative, the set $\{x \in \mathbb{R} : f(x) = c, f'(x) \neq 0\}$ has

Lebesgue measure zero [20, Theorem A.6.3]. Thus, for almost every t in the set where $B(t)=1$, we have $B'(t)=0$, and (2.22) implies $K'(t)=\langle h^s,\nu_t\rangle$. Next, by (2.11) and (2.14), if $B(t)<1$, then $Q(t)=0$. For almost every t when $Q(t)=0$, it follows that $Q'(t)=0$ and hence by (2.22) that $K'(t)=\lambda$. We have thus addressed both cases of (2.21). □

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 [27, Lemma 3.4], we will need the following notation: for any $t\in[0,\infty)$, define

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

Lemma 2.7 (Lemma 3.4 of [27]). *Suppose $\lambda\geq 0$ and Assumption 2.1 holds. Suppose $(X,\nu,\eta)=\{(X(u),\nu_u,\eta_u),\ u\geq 0\}\in\mathcal{D}_{\mathfrak{S}}[0,\infty)$ solves the fluid equations with arrival rate λ 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 λ 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 [27], 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.

2.3. Invariant states and invariant distributions of the fluid equations.

Let ν_* and η_* be Borel probability measures on $[0,\infty)$ defined as follows:

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

$$\eta_*[0,x) := \int_0^x \bar{G}^r(y)\,dy, \qquad x\in[0,H^r). \tag{2.24}$$

Note that ν_* and η_* are well defined due to Assumption 2.1. Also, recall that in the introduction, the density of the measure ν_* was denoted by f_* and thus (2.23) shows that $f_*=\bar{G}^s$. For $\lambda\geq 1$, define the set $\mathcal{X}_\lambda$ as

$$\mathcal{X}_\lambda := \left\{x\in[1,\infty): G^r\left(\left(F^{\lambda\eta_*}\right)^{-1}\left((x-1)^+\right)\right)=\frac{\lambda-1}{\lambda}\right\}, \tag{2.25}$$

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 (2.24), the map $x\to\eta_*[0,x)$ is continuous and strictly increasing on $[0,H^r)$, and therefore $(F^{\lambda\eta_*})^{-1}$ is continuous and strictly increasing on $[0,\lambda/\theta^r)$ for each $\lambda>0$. Since G^r is also continuous, we have $\mathcal{X}_\lambda=[x_l^\lambda,x_r^\lambda]$ is nonempty for each $\lambda\geq 1$. For $0\leq\lambda<1$, define $\mathcal{X}_\lambda:=\{\lambda\}$. For $\lambda\geq 0$, let $\mathcal{I}_\lambda$ be the invariant manifold for the fluid equations, defined to be the collection of invariant states of the fluid equations, and given by

$$\mathcal{I}_\lambda := \{(x_*,(\lambda\wedge 1)\nu_*,\lambda\eta_*)\in\mathfrak{S}: x_*\in\mathcal{X}_\lambda\}. \tag{2.26}$$

Our study of the critical and supercritical regimes will be carried out under the following additional assumption on the invariant manifold.

Assumption 2.8. Suppose $\lambda\geq 0$. 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_*^λ is the unique element of $\mathcal{X}_\lambda$ when $\lambda\geq 1$.

Note that Assumption 2.8 imposes a nontrivial restriction only when $\lambda \geq 1$. As stated in Lemma 3.1 of [27], a sufficient condition for Assumption 2.8 to hold when $\lambda \geq 1$ is for the equation $G^r(x) = (\lambda - 1)/\lambda$ to have a unique solution.

Whereas Assumption 2.8 guarantees a unique invariant state 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 invariant distributions, defined below. We first introduce the notion of a solution to the fluid equations when the input data is random.

DEFINITION 2.9. *Given $\lambda \geq 0$ and 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 λ 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 λ and initial condition $(X(0, \omega), \nu_0(\omega), \eta_0(\omega))$.*

DEFINITION 2.10. *For $\lambda \geq 0$, a probability measure μ on $\mathfrak{S}$ is said to be an invariant distribution of the fluid equations with arrival rate λ (sometimes abbreviated "invariant distribution for λ") if given any $\mathfrak{S}$-valued random element $(\tilde{X}, \tilde{\nu}, \tilde{\eta})$ whose law is μ, there exists a solution (X, ν, η) to the fluid equations with arrival rate λ and 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 μ.*

Remark 2.11. Fix $\lambda \geq 0$. Under our assumptions, an invariant distribution always exists. Indeed, it follows from Theorem 5.5 of [27] that the set $\mathcal{I}_\lambda$ in (2.26) describes the collection of invariant states of the fluid equations. Since for $\lambda \geq 0$, $\mathcal{X}_\lambda$ is always nonempty, an immediate consequence is that for any $z \in \mathcal{I}_\lambda$, the measure δ_z is an invariant distribution for λ. Moreover, under Assumption 2.8, $\delta_{z^\lambda_*}$ is the only invariant distribution 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 invariant distribution for λ. As shown in Proposition 4.16 below, a sufficient condition for this to hold is that any solution (X, ν, η) to the fluid equations with arrival rate λ 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) $\varepsilon^s := \operatorname{ess\,inf}_{x \geq 0} h^s(x) > 0$ and $c^s := \operatorname{ess\,sup}_{x \geq 0} h^s(x) < \infty$.

(2) The function h^s is nonincreasing.[1]

Among common probability distributions often used to model service time duration, we note that any phase type distribution satisfies Assumption 3.1(1). For the Weibull distribution $g^s(x) = \gamma x^{\gamma - 1} e^{-x^\gamma}$, the hazard rate is given by $\gamma x^{\gamma - 1}$. This function is decreasing for $0 < \gamma < 1$, providing an example for Assumption 3.1(2). 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 $H^s = \infty$.

[1] Assumption 3.1(2) should be understood in the sense of Remark 2.2, namely h^s is a.e. equal to a nonincreasing function from $[0, H^s)$ to $\mathbb{R}_+$.

The two distinct conditions imposed on the service distribution in Assumption 3.1 are adapted to two distinct techniques under which the main result is proved in this paper. Assumption 3.1(1) is required for the approach that builds on relative entropy estimates. On the other hand, the approach based on renewal equations uses Assumption 3.1(2). As mentioned in section 1.4, it would be interesting to identify a technique that would allow for a unified treatment of both cases and possibly further generalize the class of service distributions that can be handled. However, it is worth noting that an $\mathbb{L}_\infty$-bound on the coefficients akin to that implied by Assumption 3.1 is also commonly imposed in the large-time convergence analysis of age-structured equations discussed in the introduction (see, e.g., [39] or Theorem 2 of [41]).

THEOREM 3.2. *Suppose* $\lambda \geq 0$, *Assumption* 2.1 *holds, and* ν_* *and* η_* *are as defined in* (2.23) *and* (2.24), *respectively. Also, suppose* (X, ν, η) *solves the fluid equations with arrival rate* λ *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 statements are true:*

(1) *When* $0 \leq \lambda < 1$, *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 invariant distribution of the fluid equations with arrival rate* λ.

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

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

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

and

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

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_*^λ *as defined therein), then*

$$X(t) \to x_*^\lambda \quad \textit{as } t \to \infty, \tag{3.3}$$

and $\delta_{z_*^\lambda}$ *is the unique invariant distribution of the fluid equations with arrival rate* λ.

(3) *If* $\lambda = 1$ *and Assumption* 3.1(2) *is satisfied, then* $\eta_t \Rightarrow \eta_*$ *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 invariant distribution of the fluid equations with arrival rate* 1.

Remark 3.3. Regarding part (3) of the above result, note that in the critical regime $\lambda = 1$, Assumption 2.8 holds if and only if $G^r(x) > 0$ for all $x > 0$, that is, reneging within time x of arrival has positive probability for all $x > 0$. This precludes the existence of an invariant state with positive queue mass. Also, note that in this regime only uniqueness of the invariant distribution and, as $t \to \infty$, only the convergence $\langle \mathbf{1}, \nu_t \rangle \to \langle \mathbf{1}, \nu_* \rangle = 1$ are established when h^s is nonincreasing, and not the convergence $\nu_t \Rightarrow \nu_*$. In the absence of reneging the latter convergence was established when $\lambda = 1$ in Theorem 3.9(2) of [28] whenever the variance of the service distribution is finite. However, in addition to renewal estimates, the proof of this result relied on a comparison result (see Proposition 6.1(3) of [28]) that exploits certain monotonicity properties of the dynamics. Since the latter fails in the presence of reneging, it remains an open problem to extend the results in the critical regime to a broader class of service distributions. However, as elaborated below in Remark 3.4, the restricted convergence $\langle \mathbf{1}, \nu_t \rangle \to \langle \mathbf{1}, \nu_* \rangle = 1$ as $t \to \infty$ suffices for the main application of the theorem.

Remark 3.4. 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 completes (for a class of service distributions) the missing justification in the proof of the convergence result stated in Theorem 3.3 of [27].

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 scaled measure-valued N-system state process described in section 2.1 (the existence of the stationary distribution follows from Theorem 7.1 of [27]). Also, let $\bar{Z}^N := (\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)$ represent the fluid-scaled state of the N-system with initial condition $\bar{Z}^N(0) = \bar{Z}_*^N$. Then, under the assumption that $\bar{E}^N$ converges weakly to E^λ for some $\lambda \geq 0$, tightness of the sequence $\{\bar{Z}_*^N\}_{N\in\mathbb{N}}$ was established in Theorem 6.2 of [27]. Let $\bar{Z}_* = (\bar{X}_*, \bar{\nu}_*, \bar{\eta}_*)$ denote any subsequential limit of $\{\bar{Z}_*^N\}_{N\in\mathbb{N}}$. We claim that then (the law of) $\bar{Z}_*$ must be an invariant distribution of the fluid equations with arrival rate λ. To see why the claim is true, we invoke the fluid limit theorem established in Theorem 3.6 of [26] 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 (as $N \to \infty$) to $Z(t)$, where $Z := (X, \nu, \eta)$ solves the fluid equations with arrival rate λ 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 the law of $\bar{Z}_*$ is an invariant distribution.

In the proof of Theorem 3.3 in section 6.2 of [27], it was assumed without justification that $\bar{Z}_*$ is deterministic (i.e., almost surely $\bar{Z}_* = \bar{z}_*$ for some $\bar{z}_* \in \mathfrak{S}$), 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_*^λ is the weak limit of the original stationary sequence $(\bar{Z}_*^N)_{N\in\mathbb{N}}$. However, one cannot assume a priori that $\bar{Z}_*$ is deterministic, and, as argued above, one only knows that the law of any subsequential limit is an invariant *distribution,* not necessarily concentrated on an invariant state. To make this argument complete, which was one of the important motivations of this paper, one needs to show that there is precisely one invariant distribution, namely the one concentrated at the unique invariant state z_*^λ of the fluid equations with arrival rate λ. Theorem 3.2 does precisely this, imposing the condition that the class of service distributions satisfy Assumption 3.1 when $\lambda \geq 1$, thus closing the gap in the proof of the convergence result in [27] (for service distributions in that class). However, this still leaves open the question of whether this result remains true for a larger class of service distributions, in particular the entire class considered in [27].

Remark 3.5. Further, a related ancillary goal of this work is to determine whether the $N \to \infty$ and $t \to \infty$ limits commute under general convergence conditions on the initial states (essentially Assumption 3.1 of [26]), as illustrated in the diagram in Figure 1. Referring to the same notation as used in Remark 3.4, 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 [27] under some additional conditions on the service and reneging distributions (see Assumption 7.1 therein). On the other hand, as already mentioned in Remark 3.4, the left vertical arrow follows from the fluid limit theorem, Theorem 3.6 of [26] (under suitable convergence assumptions on the initial data).

$$\begin{tikzcd}
\bar{Z}^N(t)=(\bar{X}^N(t),\bar{\nu}^N_t,\bar{\eta}^N_t) \arrow[r, Rightarrow, "\text{Thm 7.1 of [27]}"] \arrow[d, Rightarrow, "\text{Thm 3.6 of [26]}"] & \bar{Z}^N_*=(\bar{X}^N_*,\bar{\nu}^N_*,\bar{\eta}^N_*) \arrow[d, Rightarrow, "\text{Thm 6.2 of [27] and }\textbf{Thm 3.2}"] \\
Z(t)=(X(t),\nu_t,\eta_t) \arrow[r, "\textbf{Thm 3.2}"] & z_*=(x_*,(\lambda\wedge 1)\nu_*,\lambda\eta_*)
\end{tikzcd}$$

FIG. 1. *Interchange of limits diagram.*

Along with the tightness of $(\bar{Z}^N_*)_{N\in\mathbb{N}}$ established in [27], 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 3.4) 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 convergence 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 [28] (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* (2.10). *Then $\eta_t\Rightarrow\lambda\eta_*$ as $t\to\infty$.*

Proof. Fix $\psi\in\mathcal{C}_b(\mathbb{R}_+)$. In view of (2.19), the boundedness of ψ, the finiteness of the measure η_0, the dominated convergence theorem, and the fact that $\bar{G}^r(x+t)/\bar{G}^r(x)\to 0$ as $t\to\infty$, for every $x\in[0,H^r)$, together imply that the first term on the right-hand side of (2.19) vanishes. On the other hand, since the mean patience time $\int_0^\infty\bar{G}^r(u)du$ is finite, the dominated convergence theorem shows that the last term on the right-hand side of (2.19) converges to $\langle\psi,\lambda\eta_*\rangle$. This concludes the proof that $\eta_t\Rightarrow\lambda\eta_*$ as $t\to\infty$. ∎

4.1. Proof of Theorem 3.2(1).

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,ν,η) is a solution to the fluid equations for arrival rate λ and initial condition $(X(0),\nu_0,\eta_0)$, and let (D,K,R,S,Q,B) be the corresponding auxiliary processes.

The weak convergence of η_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 (2.7), setting $\psi=h^s$ in (2.20), interchanging the order of integration, and using integration by parts and the fact $G^s(0+)=0$, we obtain for $t\geq 0$,

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

Substituting this in (2.12), using (2.17) and (2.13), and performing repeated integration by parts, we obtain for $t \geq 0$,

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

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

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

We now make use of the following simple observation.

LEMMA 4.2. $\limsup_{t\to\infty} \int_0^t Q(u) g^s(t-u)du \leq \limsup_{t\to\infty} Q(t)$.

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

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

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

Continuing with the proof of Theorem 3.2(1), taking the limit supremum in (4.3), and using Lemma 4.2, the identity $\int_0^\infty \bar{G}^s(u)du = 1$ 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 (2.14), we obtain

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

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 (2.16), 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 λ and initial condition $(X(T'), \nu_{T'}, \eta_{T'})$ and hence, (2.20) holds with ν and K replaced with $\nu^{[T']}$ and $K^{[T']}$, respectively. Since $\nu^{T'}(t) = \nu_{T'+t}$ and by (2.16), (2.13), and (2.17), $Q(T' + \cdot) \equiv 0$, $R^{[T']}(\cdot) \equiv 0$, and $K^{T'}(t) = K(T'+t) - K(T') = \lambda t$, $t \geq 0$, this implies that for every $\psi \in \mathcal{C}_b[0, H^s)$,

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

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 $\lambda \int_{[0,H^s)} \psi(x)\bar{G}^s(x)dx$ (where, for the latter, recall that $\int_0^\infty \bar{G}^s(u)du = 1$ from Assumption 2.1). Recalling that $\nu_*(dx) = f_*(x)dx = \bar{G}^s(x)dx$, 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 (2.11) 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 λ. 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 π of an invariant distribution 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 (2.11) that $X(0) = \lambda$ almost surely, thus proving that $\pi = \delta_{z_*^\lambda}$ with $z_*^\lambda = (\lambda, \lambda\eta_*, \lambda\nu_*)$. This completes the proof of Theorem 3.2(1).

4.2. Proof of Theorem 3.2(2a) 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). Suppose Assumption 2.1 is satisfied and Assumption 3.1(1) holds (with associated constants $\varepsilon^s > 0$ and $0 < c^s < \infty$). Recall that $f_*(x) = \bar{G}^s(x)$ is the density of ν_*. 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, ν, η) is the associated solution to the fluid equations. We will establish convergence, as $t \to \infty$, of the fluid age measure ν_t described in (2.8) by appealing to properties of an extended relative entropy functional, which we now introduce.

Recall that $\mathcal{M}_F(E)$ and $\mathcal{P}(E)$ denote the spaces of finite nonnegative Borel measures and Borel probability measures, respectively, on a measurable space E, and define the functional $R(\cdot\|\cdot) : \mathcal{M}_F(E) \times \mathcal{P}(E) \mapsto (-\infty, \infty]$ by

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

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 subprobability 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 $P=Q$.

As alluded to in the introduction, the idea of proving Theorem 3.2(2)(a) is to use the generalized relative entropy function $R(\cdot\|\nu_*)$ in a manner reminiscent of a Lyapunov function. In section 4.2.1 we state some estimates related to this function. Then, in section 4.2.2, we carry out in greater detail the formal calculations described in sections 1.1 and 1.2. These calculations aim only at providing intuition into why $R(\cdot\|\nu_*)$ is a candidate Lyapunov function for the problem at hand, and making some of the calculations carried out later in the proof more transparent. However, this section is purely motivational, and the reader could also skip directly to the more involved rigorous proof given in section 4.2.3.

4.2.1. Estimates related to the extended relative entropy functional. The proof of Theorem 3.2(2) will make use of two properties of extended relative entropy which we now state.

Lemma 4.4. *Suppose P and Q are finite nonnegative Borel measures on $\mathbb{R}_+$, equipped with the Borel σ-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., p. 44 of [15]) 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 Borel measurable function that satisfies $\int_0^\infty f(x)dx \leq 1$ and suppose that $z_f := \int_0^\infty h^s(x)f(x)dx < \infty$ and μ^f is the measure with density f. Then*

$$\begin{aligned}\int_0^\infty h^s(x)f(x)\log\frac{f(x)}{f_*(x)}dx &\geq z_f\log z_f \\ &\quad + \varepsilon^s\int_0^\infty f(x)\log\frac{f(x)}{f_*(x)}dx = z_f\log z_f + \varepsilon^s R(\mu^f\|\nu_*).\end{aligned} \tag{4.8}$$

4.2.2. A formal calculation. Although Theorem 3.2(2)(a) is concerned with the supercritical case, the calculations here are valid for any $\lambda \geq 0$. Observe that (2.8) expresses $(\nu_t)_{t\geq 0}$ as a weak solution to a transport equation (with data K). Now, for the purposes of this formal calculation only, suppose that ν_0 has a density, denoted by f_0, and for each $t>0$, suppose the measure ν_t has a sufficiently smooth density, denoted by $f(x,t)$, $x \geq 0$. Then by (2.18) and (2.7), $\langle \mathbf{1}, \nu_t \rangle = \int_0^\infty f(x,t)dx$ and $\langle h^s, \nu_t \rangle = \int_0^\infty h^s(x) f(x,t)dx$, the transport equation could be formally rewritten as the PDE specified in (1.1)–(1.3), which we repeat here for convenience:

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

with the boundary condition $f(0,t) = k(t)$, which by (2.21) takes the form

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

and the initial condition

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

As mentioned in the introduction, (4.9)–(4.11) may in general have multiple solutions, and a precise analysis must consider the dynamics of ν together with the other components. Proceeding with purely formal calculations to gain intuition, note that $f_*(x) = e^{-J(x)}$, where $J(x) := \int_0^x h^s(y)dy < \infty$ for every $x > 0$. For $t > 0$, define

$$\hat{\Delta}(t) := R(\nu_t \| \nu_*) = \int_0^\infty f(x,t) \log \frac{f(x,t)}{f_*(x)} dx = \int_0^\infty f(x,t)(\log f(x,t) + J(x))dx.$$

(We use $\hat{\Delta}(t)$ to distinguish it from another function, $\Delta(t)$, that we use in the actual proof that is defined slightly differently from $\hat{\Delta}(t)$.) Taking derivatives of both sides of the last equation with respect to t, and using (4.9), we see that

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

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

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

On combining the last two equations, and recalling that $J(x) = -\log f_*(x)$, we obtain

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

Note that $\int_0^\infty f(x,t)dx = \langle \mathbf{1}, \nu_t\rangle \le 1$ and that for almost every $t \in [0,\infty)$, (2.7) implies that $z_{f(\cdot,t)} = \int_0^\infty h^s(x) f(x,t)dx < \infty$. Hence Lemma 4.5 is applicable with $f = f(\cdot,t)$ and we obtain

$$\begin{aligned}\int_0^\infty h^s(x) f(x,t) \log \frac{f(x,t)}{f_*(x)} dx &\ge \left(\int_0^\infty h^s(x) f(x,t)dx\right) \log \left(\int_0^\infty h^s(x) f(x,t)dx\right) \\ &\quad + \varepsilon^s \int_0^\infty f(x,t) \log \frac{f(x,t)}{f_*(x)} dx \\ &= z_{f(\cdot,t)} \log z_{f(\cdot,t)} + \varepsilon^s \hat{\Delta}(t).\end{aligned}$$

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

$$(4.12) \qquad \frac{d}{dt}\hat{\Delta}(t) \le \begin{cases} -\varepsilon^s \hat{\Delta}(t) + \lambda \log \lambda - z_{f(\cdot,t)} \log z_{f(\cdot,t)} & \text{if } \int_0^\infty f(x,t)dx < 1, \\ -\varepsilon^s \hat{\Delta}(t) & \text{if } \int_0^\infty f(x,t)dx = 1. \end{cases}$$

As already observed in the introduction, this estimate does not directly imply the convergence of $\hat{\Delta}(t)$ to zero. However, the fact that it takes the form $\frac{d\hat{\Delta}}{dt}(t) \le -\varepsilon^s \hat{\Delta}(t)$ in the case $\int_0^\infty f(x,t)dx = 1$ 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, $\int_0^\infty f(x,t)dx = 1$. However, converting this into a fully rigorous argument is rather nontrivial. The argument provided in the next section is considerably more involved and copes with the more complicated structure of the estimate in the case $\int_0^\infty f(x,t)dx < 1$, as well as the fact that $\hat{\Delta}(t)$ can go negative.

4.2.3. Proof of Theorem 3.2(2)(a). First, note that the limit $\eta_t \Rightarrow \lambda\eta_*$ in (3.1) holds for any $\lambda \ge 0$ from Lemma 4.1. To establish the remaining limits, we begin with the representation for the age measure ν_t given in (2.20), which shows that $\nu_t = \theta_t + \mu_t$, where $\theta_t, \mu_t \in \mathcal{M}_F[0,\infty)$, are defined by

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

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

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

where we recall that k, defined in (2.21), is a.e. equal to the derivative K' of the auxiliary process K defined in (2.9) of the fluid equations.

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

$$(4.15) \qquad \begin{aligned} d_{\mathrm{TV}}(\mu_t, \nu_*) = \int_0^\infty \left|\tilde{f}(x,t) - f_*(x)\right| dx &\le |\langle \mathbf{1}, \mu_t\rangle - 1| \\ &\quad + \left(2\langle \mathbf{1}, \mu_t\rangle^{-1} |\Delta(t)| + 2|\log\langle \mathbf{1}, \mu_t\rangle|\right)^{1/2},\end{aligned}$$

where for $t \geq 0$,

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

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

with the last equality using the fact that $k(t-x) = \tilde{f}(x,t)/f_*(x)$ for $x \in [0,t]$, due to (4.14). Since the expression in (2.21) and Assumption 3.1(1) show that k is strictly positive and bounded above by $\lambda \vee c^s = \lambda \vee \sup_{x\in[0,\infty)} h^s(x)$, $\Delta(t)$ is well defined and finite.

Remark 4.6. Due to the pointwise convergence $\frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \to 0$ as $t \to \infty$ for each $x > 0$, the dominated convergence theorem shows that $\langle \mathbf{1}, \theta_t \rangle$, the total mass of θ_t, converges to zero as $t \to \infty$. Hence, θ_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 $\Delta(t) \to 0$ as $t \to \infty$.

Our main goal in this section is to establish these limits in the supercritical regime $\lambda > 1$.

PROPOSITION 4.7. *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,*

$$\langle \mathbf{1}, \mu_t \rangle \to 1 \qquad \textit{and} \qquad \Delta(t) \to 0, \qquad \textit{as } t \to \infty, \tag{4.18}$$

and also $\nu_t \Rightarrow \nu_*$ *and* $\langle h^s, \nu_t \rangle \to 1$ *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 4.7 in Step 4.

Step 1. We start with simple bounds on quantities associated with the measure-valued function θ_t defined in (4.13). Recall the convention $0 \log 0 = 0$.

LEMMA 4.8. *We have the following:* $\sup_t \langle h^s, \theta_t \rangle \leq c^s$, $\int_0^\infty \langle h^s, \theta_t \rangle dt < \infty$, *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 $t \geq 0$, $\langle h^s, \theta_t \rangle \leq c^s \langle \mathbf{1}, \nu_0 \rangle \leq c^s$. This proves the first bound.

With a view to establishing the remaining two bounds we first prove a refinement of the above bound; specifically, we show that $\langle h^s, \theta_t \rangle \leq c^s e^{-\varepsilon^s t}$ for all $t \geq 0$. To this end, first use the relation $\bar{G}^s(y) = e^{-\int_0^y h^s(u)du}$ and the definition of ε^s as the essential infimum of h^s (see Assumption 3.1(1)) to conclude that for all $x \geq 0$ and $t \geq 0$, $\bar{G}^s(x+t) \leq \bar{G}^s(x) e^{-\varepsilon^s t}$. Substituting this into (4.19) and invoking the definition of c^s as the supremum of h^s (see Assumption 3.1(1)), this yields

$$\langle h^s, \theta_t \rangle \leq e^{-\varepsilon^s t} \int_{[0,\infty)} h^s(x+t) \nu_0(dx) \leq c^s e^{-\varepsilon^s t} \langle \mathbf{1}, \nu_0 \rangle \leq c^s e^{-\varepsilon^s t}, \tag{4.20}$$

which immediately implies the second bound: $\int_0^\infty \langle h^s, \theta_t \rangle dt < \infty$.

To prove the last assertion of the lemma, note that on $(0,e^{-1}]$, the function $x\mapsto -x\log x$ is increasing and nonnegative. Let $t_0>0$ be such that $c^s e^{-\varepsilon^s t_0}<e^{-1}$. Then using (4.20) for $t>t_0$, we obtain

$$|\langle h^s,\theta_t\rangle\log\langle h^s,\theta_t\rangle| = -\langle h^s,\theta_t\rangle\log\langle h^s,\theta_t\rangle \leq -c^s e^{-\varepsilon^s t}\log(c^s e^{-\varepsilon^s t}) = c^s(\varepsilon^s t-\log c^s)e^{-\varepsilon^s t},$$

which in turn implies $\int_{t_0}^{\infty}|\langle h^s,\theta_t\rangle\log\langle h^s,\theta_t\rangle|dt<\infty$. On the other hand, by the first bound of the lemma,

$$\int_0^{t_0}|\langle h^s,\theta_t\rangle\log\langle h^s,\theta_t\rangle|dt \leq t_0\sup_{t\in[0,t_0]}|\langle h^s,\theta_t\rangle\log\langle h^s,\theta_t\rangle| \leq t_0\sup_{[0,c^s]}|x\log x|.$$

But, this is also finite since the function $[0,\infty)\ni x\mapsto x\log x$, which takes the value 0 when $x=0$, is continuous, and hence bounded on $[0,t_0]$. When combined, the last two statements prove the last assertion and complete the proof of the lemma. □

Step 2. We now establish our main estimate on $\Delta(t)$ in Corollary 4.10, building off preliminary estimates that we first obtain in Lemma 4.9. 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 $[0,c^s]$, i.e., $m(\delta)=\sup\{|x\log x-y\log y| : 0\leq x\leq y\leq c^s\wedge(x+\delta)\}$. On $[0,e^{-1}]$ this function is decreasing. Now, for $0\leq x<y\leq e^{-1}$, applying the inequality $p\log p+(1-p)\log(1-p)\leq 0$ (which holds for $p\in[0,1]$) with $p=x/y$, we see that

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

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

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

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

$$m(\delta)\leq|\delta\log\delta|+c_1\delta,\qquad \delta\in[0,c^s]. \tag{4.21}$$

LEMMA 4.9. *For $t\geq 0$, define $\Upsilon_t:=m(\langle h^s,\theta_t\rangle)$, where θ_t is defined by* (4.13)*. If (t_1,t_2) is a busy interval, then*

$$\Delta(t)\leq\Delta(t_1)e^{-\varepsilon^s(t-t_1)}+\int_{t_1}^{t}\Upsilon_u du,\quad t\in(t_1,t_2). \tag{4.22}$$

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

$$\Delta(t_2)\leq\Delta(t_1)+\int_{t_1}^{t_2}\Upsilon_u du, \tag{4.23}$$

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_Δ and $c_{\rm lip}$ such that $\sup_t|\Delta(t)|\leq c_\Delta$ and for any $0\leq u<t<\infty$, $|\Delta(t)-\Delta(u)|\leq c_{\rm lip}|t-u|$, showing that the function $t\to\Delta(t)$ is bounded and 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 Δ is differentiable (although not continuously differentiable) with derivative

$$\begin{aligned}\frac{d\Delta}{dt}(t) &= k(t)\log k(t) - \int_0^t g^s(t-x)k(x)\log k(x)dx\\ &= k(t)\log k(t) - \int_0^t g^s(x)k(t-x)\log k(t-x)dx. \end{aligned}\tag{4.25}$$

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) and the convention $0\log 0 = 0$, and then applying Lemma 4.5, with $f(x)$ replaced with $\tilde f(x,t)$ (since $\int_0^\infty \tilde f(x,t)dx \le \langle 1, \nu_t\rangle \le 1$), we obtain

$$\begin{aligned}\frac{d\Delta(t)}{dt} &= k(t)\log k(t) - \int_0^\infty h^s(x)\tilde f(x,t)\log\frac{\tilde f(x,t)}{f_*(x)}dx\\ &\le k(t)\log k(t) - z_{\tilde f(\cdot,t)}\log z_{\tilde f(\cdot,t)} - \varepsilon^s\Delta(t),\end{aligned}\tag{4.26}$$

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 (2.21) and Assumption 3.1(1), for $t \in (t_1,t_2)$, $c^s \ge k(t) = \langle h^s, \nu_t\rangle = \langle h^s, \mu_t\rangle + \langle h^s, \theta_t\rangle$, 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 $[0, c^s]$, 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),

$$\frac{d\Delta}{dt}(t) \le \Upsilon_t - \varepsilon^s\Delta(t), \quad t \in (t_1,t_2).\tag{4.27}$$

Now, let $\tilde\Delta$ denote the solution to the differential equation $\frac{d\tilde\Delta}{dt}(t) = \Upsilon_t - \varepsilon^s\tilde\Delta(t)$ with the same initial condition as Δ, namely $\tilde\Delta(t_1) = \Delta(t_1)$. Then $\tilde\Delta$ can be solved explicitly:

$$\tilde\Delta(t)=(\tilde\Delta(t_1))^{-\varepsilon^s(t-t_1)} + \int_{t_1}^t e^{-\varepsilon^s(u-t_1)}\Upsilon_{t-u}du \le \Delta(t_1)e^{-\varepsilon^s(t-t_1)} + \int_{t_1}^t \Upsilon_u du,\ \ t\in(t_1,t_2).$$

A standard comparison theorem for ordinary differential equations yields $\Delta(t) \le \tilde\Delta(t)$ for $t \in (t_1,t_2)$. This proves (4.22).

Next, consider an excursion interval (t_1,t_2). Then (2.21) implies that $k(t) = \lambda$ for $t \in (t_1,t_2)$. Moreover, it is not hard to see that the fluid age equation (2.8) 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, $D(t_2) - D(t_1) = K(t_2) - K(t_1) = \lambda(t_2 - t_1)$ and it follows that

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

Thus, the convexity of $x\log x$ and Jensen's inequality together imply

$$\lambda\log\lambda \le \frac{1}{t_2-t_1}\int_{t_1}^{t_2}\langle h^s,\nu_t\rangle\log\langle h^s,\nu_t\rangle dt. \tag{4.29}$$

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 (4.29) we have

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

We now turn to the last assertion of the lemma. The bound $0\le k(t)\le c^s\vee\lambda$ (by continuity of the function $\mathbb{R}_+\ni x\mapsto x\log x$) implies that $|k(t)\log k(t)|\le c_2$ for some finite constant c_2. The boundedness of $t\to\Delta(t)$ 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 $\frac{d\Delta}{dt}(t)$ is bounded and hence that $t\mapsto\Delta(t)$ is globally Lipschitz on $[0,\infty)$. □

As a corollary, we obtain our main estimate on $\Delta(t)$. For $t>0$, define

$$L(t):=\int_0^t 1_{\{B(u)=1\}}du,\quad t>0,\qquad\text{and}\qquad \mathcal{B}:=\{t>0:B(t)=1\}. \tag{4.30}$$

COROLLARY 4.10. *For every $\bar t\ge 0$ and $t>\bar t$, $t\in\mathcal{B}$,*

$$\Delta(t)\le c_\Delta e^{-\varepsilon^s(L(t)-L(\bar t))}+\int_{\bar t}^t\Upsilon_\tau d\tau, \tag{4.31}$$

where c_Δ is the constant from Lemma 4.9*.*

Proof. Fix $\bar t\ge 0$ and $t>\bar t$, $t\in\mathcal{B}$. Denote $t_0:=\inf\{u\ge\bar t:B_u=1\}$. 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 $\mathcal{O}:=\{u\in(\bar t,t):B_u<1\}$ 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'), $n=1,\dots,I_u$, in such a way that $\bar t\le t_0\le t_1<t_2<t_3<\cdots<t_{I_u}\le t$. Let c_{lip} denote the (global) Lipschitz constant of $t\mapsto\Delta(t)$, which exists by Lemma 4.9. We now show by induction that, for $n=1,2,\dots,I_u$,

$$\Delta(t_n)\le\Delta(t_1)e^{-\varepsilon^s\sum_{i=1}^{n-1}(L(t_i')-L(t_i))}+\sum_{i=1}^{n-1}\int_{t_i}^{t_i'}\Upsilon_\tau d\tau+c_{\text{lip}}\sum_{i=1}^{n-1}(t_{i+1}-t_i'), \tag{4.32}$$

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

Base case: For $n=1$, (4.32) reduces to the trivial inequality $\Delta(t_1)\le\Delta(t_1)$ and thus is satisfied.

Induction step: Assuming (4.32) holds for an arbitrary $n\in\{1,\dots,I_u-1\}$, we show it holds for $n+1$. From (4.22) and (4.23) of Lemma 4.9, 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

$$\Delta(t'_n)\le\Delta(t_n)e^{-\varepsilon^s(L(t'_n)-L(t_n))}+\int_{t_n}^{t'_n}\Upsilon_\tau d\tau.$$

Using this estimate, the Lipschitz continuity of Δ established in Lemma 4.9, and the induction hypothesis, it follows that

$$\begin{aligned}\Delta(t_{n+1})&\le\Delta(t'_n)+c_{\rm lip}(t_{n+1}-t'_n)\\&\le\Delta(t_n)e^{-\varepsilon^s(L(t'_n)-L(t_n))}+\int_{t_n}^{t'_n}\Upsilon_\tau d\tau+c_{\rm lip}(t_{n+1}-t'_n)\\&\le\Big(\Delta(t_1)e^{-\varepsilon^s\sum_{i=1}^{n-1}(L(t'_i)-L(t_i))}+\sum_{i=1}^{n-1}\int_{t_i}^{t'_i}\Upsilon_\tau d\tau\\&\qquad+c_{\rm lip}\sum_{i=1}^{n-1}(t_{i+1}-t'_i)\Big)e^{-\varepsilon^s(L(t'_n)-L(t_n))}\\&\qquad\quad+\int_{t_n}^{t'_n}\Upsilon_\tau d\tau+c_{\rm lip}(t_{n+1}-t'_n)\\&\le\Delta(t_1)e^{-\varepsilon^s\sum_{i=1}^{n}(L(t'_i)-L(t_i))}+\sum_{i=1}^{n}\int_{t_i}^{t'_i}\Upsilon_\tau d\tau+c_{\rm lip}\sum_{i=1}^{n}(t_{i+1}-t'_i).\end{aligned}$$

This proves (4.32) by induction.

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

$$\begin{aligned}L(t)-L(t_0)-\sum_{i=1}^{I_u-1}(L(t'_i)-L(t_i))&=(L(t_1)-L(t_0))\\&\quad+\sum_{i=1}^{I_u-1}(L(t_{i+1})-L(t'_i))+(L(t)-L(t_{I_u}))\le|\mathcal{T}_u|.\end{aligned}$$

Hence, on applying (4.32) with $n=I_u$ and noting that Υ is nonnegative and the last term on the right-hand side is bounded above by $c_{\rm lip}|\mathcal{T}_u|$, we obtain

$$\Delta(t_{I_u})\le c_\Delta e^{-\varepsilon^s(L(t)-L(t_0)-|\mathcal{T}_u|)}+\int_{t_0}^{t}\Upsilon_\tau d\tau+c_{\rm lip}|\mathcal{T}_u|,\quad u>0.$$

Then noting that since $(t_{I_u},t)\subset\mathcal{T}_u$ for all $u>0$, it follows that $\lim_{u\to0+}t_{I_u}=t$. Finally, the result follows on sending $u\to0$ since Δ is continuous and by definition $\bar t\le t_0$ and $L(t_0)=L(\bar t)$. ∎

Step 3. Now, assuming $\lambda>1$, we 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. (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), $\langle h^s, \nu_t\rangle \leq c^s$, for all $t > 0$. However, (4.13), (4.14), and (2.21) together imply

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

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 4.7.

Proof of Proposition 4.7. Fix $\lambda > 1$. We first claim that to establish (4.18), it suffices to show that $B(t) = 1$ for all sufficiently large t. Recalling that $\Upsilon_t = m(\langle h^s, \theta_t\rangle)$ is integrable on $[0,\infty)$ by Lemma 4.8 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 a sequence in $\mathcal{B}$ and then $\bar{t} \to \infty$ in (4.31) of Corollary 4.10, to obtain $\limsup_{t\to\infty,\ t\in\mathcal{B}} \Delta(t) \leq 0$. We cannot directly deduce from this that the limit of $\Delta(t)$ along $\mathcal{B}$ is zero, since $\Delta(t) = R(\mu_t\|\nu_*)$ 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 $\Delta(t) = R(\mu_t\|\nu_*)$, and as $t \to \infty$, $\langle \mathbf{1}, \theta_t\rangle \to 0$ by Remark 4.6, when combined with (4.6) this implies $\limsup_{t\to\infty,t\in\mathcal{B}} \Delta(t) \geq \limsup_{t\to\infty,t\in\mathcal{B}}\langle \mathbf{1}, \mu_t\rangle \ln\langle \mathbf{1}, \mu_t\rangle = 0$. Hence,

$$\lim_{t\to\infty,\ t\in\mathcal{B}} \langle \mathbf{1}, \mu_t\rangle = 1 \quad \text{and} \quad \lim_{t\to\infty,\ t\in\mathcal{B}} \Delta(t) = 0, \tag{4.33}$$

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 (4.33) 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

$$c^s\langle \mathbf{1}, \theta_t\rangle < \varepsilon_0 \quad \text{and} \quad c^s\int_0^\infty |\tilde{f}(x,t) - f_*(x)|dx < \varepsilon_0 \qquad \text{for all } t \geq T,\ t \in \mathcal{B}. \tag{4.34}$$

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 particular, the estimates in (4.34) are valid for $t=\tau$. Find $t^*>0$ so small that $G^s(t^*)<\frac{1}{4}$ and $0<t^*<T'-\tau$. For all $0\le t\le t^*$, applying Lemma 2.7 with $t=\tau$ and (2.20) with $\psi=h^s$, and using the fact that (2.21) 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 (4.34), we obtain

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

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

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

which is a contradiction. We have thus shown that $B(t)=1$ for all sufficiently large t. Together with (4.33), this proves (4.18).

To conclude the proof of the proposition, it only remains to show that $\nu_t\Rightarrow\nu_*$ and $\langle h^s,\nu_t\rangle\to 1$ as $t\to\infty$. Fix $T\in(0,\infty)$ such that $B(t)=1$ for all $t\ge T$. 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) in Corollary 4.4 of [28], K has the representation

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

In view of the representation for the fluid age measure in (2.20), the convergence $\nu_t\Rightarrow\nu_*$ is then a direct consequence of Lemma 6.2 of [28] with $\pi=\nu$. Next, to show $\langle h^s,\nu_t\rangle\to 1$ as $t\to\infty$, using Lemma 2.7 and (2.20) with $\psi(x)=h^s(x)$, and noting from (2.21) that $K'(T+u)=\langle h^s,\nu_{T+u}\rangle$ for all $u>0$, and recalling again that $g^s=\bar{G}^sh^s$, we have

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

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

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

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 nonincreasing. In this section, we assume throughout that Assumptions 2.1 and 3.1(2) hold, and we establish Theorem 3.2(2) in this case, as well as Theorem 3.2(3). In addition, for some $\lambda \geq 0$, let (X,ν,η) be the solution to the fluid equations with arrival rate λ and some initial condition $(X(0),\nu_0,\eta_0)\in\mathfrak{S}$. Also, recall from (2.18) that $B(t)=\langle \mathbf{1},\nu_t\rangle$, and define

$$M(t) := B(t) - \int_{[0,H^s)}\frac{\bar G^s(x+t)}{\bar G^s(x)}\nu_0(dx), \quad t\geq 0. \tag{4.35}$$

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

We will first establish the following key result.

Proposition 4.11. *Suppose Assumptions* 2.1 *and* 3.1(2) *hold, and* $\lambda\geq 1$. *Then we have*

$$\lim_{t\to\infty} M(t) = \lim_{t\to\infty} B(t) = 1. \tag{4.36}$$

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 that are valid for any $\lambda\geq 0$. Setting $\psi\equiv\mathbf{1}$ in (2.20) and using integration by parts, it follows that for each $t\geq 0$,

$$\begin{aligned}B(t)=\langle\mathbf{1},\nu_t\rangle &= \int_{[0,H^s)}\frac{\bar G^s(x+t)}{\bar G^s(x)}\nu_0(dx) + \int_0^t \bar G^s(t-w)dK(w)\\ &= \int_{[0,H^s)}\frac{\bar G^s(x+t)}{\bar G^s(x)}\nu_0(dx) + K(t) - \int_0^t K(w)g^s(t-w)dw,\end{aligned} \tag{4.37}$$

which when rearranged yields

$$K(t) = M(t) + \int_0^t K(t-w)g^s(w)dw. \tag{4.38}$$

Then (4.37), (4.35), and the fact that ν_t is a subprobability measure together imply that for each $t\geq 0$,

$$M(t) = \int_0^t \bar G^s(t-w)dK(w)\geq 0 \quad\text{and}\quad M(t)\leq B(t)\leq 1. \tag{4.39}$$

Together with (4.38) and the renewal theorem (see Chapter V of [7]), this implies that for each $t\geq 0$

$$K(t) = M(t) + Z(t) \qquad\text{with}\quad Z(t) := \int_0^t M(t-w)\ dU^s(w), \tag{4.40}$$

where U^s is the renewal function, $U^s(w) := \sum_{n=0}^{\infty}(G^s)^{\star n}(w), w \geq 0$, and for $n \geq 0$, $(G^s)^{\star n}$ denotes the nth convolution power of G^s with the convention that $(G^s)^{\star 0}(w) = 1$ for all $w \geq 0$. Note that $Z(t)$ represents the mass of jobs that entered service after time 0 and has completed service and departed the system by time t. Now, (2.7) and (2.9) imply that

$$D(t) = \int_0^t \langle h^s, \nu_w \rangle dw = B(0) - B(t) + K(t), \qquad t \geq 0.$$

Then by (4.40), (4.35), and (4.37), we obtain

$$\begin{aligned} D(t) &= \langle \mathbf{1}, \nu_0 \rangle - \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) + Z(t) \\ &= \int_{[0,H^s)} \frac{G^s(x+t) - G^s(x)}{\bar{G}^s(x)} \nu_0(dx) + Z(t). \end{aligned} \tag{4.41}$$

Under Assumption 3.1(2), the hazard rate function h^s is nonincreasing 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 [7] 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 (e.g., see p. 172 of [37]), the concavity of U^s implies that u^s is nonincreasing and differentiable a.e., that is,

$$(u^s)'(t) \leq 0 \quad \text{ for a.e. } t \geq 0. \tag{4.42}$$

Now, differentiation of both sides of the defining equation for Z in (4.40) yields for a.e. $t \geq 0$,

$$Z'(t) = M(t)u^s(0) + \int_0^t M(t-w)(u^s)'(w)dw. \tag{4.43}$$

On the other hand, differentiating the equation for K in (4.40) and using (2.21), one obtains, for a.e. $t \geq 0$,

$$\begin{aligned} M'(t) &= K'(t) - Z'(t) \\ &= \begin{cases} \lambda - Z'(t) & \text{if } B(t) < 1, \\ D'(t) - Z'(t) & \text{if } B(t) = 1. \end{cases} \end{aligned} \tag{4.44}$$

Next, differentiating both sides of (4.41), 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 (4.44), for a.e. $t \geq 0$,

$$\text{if } Z'(t) \leq \lambda, \text{ then } M'(t) \geq 0. \tag{4.45}$$

We now establish some auxiliary results that will be used in the proof of Proposition 4.11.

LEMMA 4.12. *Suppose $\lambda \geq 1$. Then there is no $T \in (0,\infty)$ and $c \in (0,1)$ such that $M(t) < c$ for all $t \geq T$. The same assertion also holds when M 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 $M(t) < c$ 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 (4.35), there exists $T' > T$ such that $B(t) < 1$ for all $t \geq T'$. In turn, by (2.9), it follows that $K'(t) = \lambda$, for all $t \geq T'$, and hence (4.37) and (4.35) imply that

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

As $t \to \infty$, the first term converges to zero by the dominated convergence theorem and the pointwise limit $\bar{G}^s(t-w) \to 0$. For the same reason, the second term converges to $\lim_{t\to\infty} \lambda \int_0^t \bar{G}^s(t-w)dw = \lambda \int_0^\infty \bar{G}^s(w)dw$, which is equal to λ by (2.1) of Assumption 2.1. Thus, $\lim_{t\to\infty} M(t) \geq \lambda$, which is a contradiction, thus proving the first assertion of the lemma. Since, by (4.35), $B(t) - M(t) = \int_{[0,H^s)} \frac{\bar{G}^s(x+t)}{\bar{G}^s(x)} \nu_0(dx) \to 0$ as $t \to \infty$, the same assertion holds also for B. □

Next, substituting into (4.43) the inequality (4.42), the relation $u^s(0) = g^s(0)$, and the fact that $M(t) \in [0,1]$ for each $t \geq 0$ due to (4.39), we see that

$$Z'(t) \leq M(t)u^s(0) = M(t)g^s(0) \leq g^s(0) \quad \text{for a.e. } t \geq 0. \tag{4.46}$$

We also observe that since the hazard rate function h^s is nonincreasing 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

$$\begin{aligned} \lambda_n := \frac{\lambda-\varepsilon}{g^s(0)} \left(\sum_{i=0}^n \left(1 - \frac{1}{g^s(0)}\right)^i \right) &= \frac{\lambda-\varepsilon}{g^s(0)} \frac{1 - \left(1 - \frac{1}{g^s(0)}\right)^{n+1}}{1 - \left(1 - \frac{1}{g^s(0)}\right)} \\ &= (\lambda - \varepsilon)\left(1 - \left(1 - \frac{1}{g^s(0)}\right)^{n+1}\right), \end{aligned} \tag{4.47}$$

and

$$\tau_n := \sup\{t > 0 : M(t) < \lambda_n\}. \tag{4.48}$$

If $\tau_n < \infty$, then

$$M(\tau_n + t) \geq \lambda_n \quad \text{for all } t \geq 0. \tag{4.49}$$

LEMMA 4.13. *Suppose $\lambda \geq 1$, $\varepsilon \in (0, \frac{1}{2})$ and $g^s(0) > 1 \vee (\lambda - \varepsilon)$. Then $\tau_n < \infty$ and hence* (4.49) *holds for all $n \in \mathbb{N}$ with $n < n^*$, where $n^* := \sup\{n \in \mathbb{N} \cup \{0\} : \lambda_n < 1\}$, and also for $n = n^*$, if $n^* < \infty$.*

Proof. Since $g^s(0) > 1 \vee (\lambda - \varepsilon) \geq 1$ by the assumptions of the lemma, it follows that $1 - \frac{1}{g^s(0)} \in (0,1)$ and (4.47) then implies that

$$\lambda_n \uparrow \lambda - \varepsilon \text{ as } n \to \infty. \tag{4.50}$$

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

$$M(t) < \frac{\lambda - \varepsilon}{g^s(0)} \quad \text{for all } t \in (0, \tau_0). \tag{4.51}$$

Note that (4.51) holds trivially if $\tau_0 = 0$. So, suppose $\tau_0 > 0$ and (4.51) does not hold. Then there must exist $0 < t_1 < \tau_0$ for which $M(t_1) \geq \frac{\lambda-\varepsilon}{g^s(0)}$. It follows from (4.46) and (4.45) that for a.e. $t \in (t_1, \tau_0)$, the inequality $M(t) < \frac{\lambda-\varepsilon}{g^s(0)}$ implies that $Z'(t) \leq M(t)g^s(0) < \lambda - \varepsilon < \lambda$ and hence $M'(t) \geq 0$. Since M is absolutely continuous, $M(t) \geq \frac{\lambda-\varepsilon}{g^s(0)} = \lambda_0$ for all $t \in [t_1, \tau_0)$ (see Lemma B.1). This contradicts the definition of τ_0, and thus (4.51) holds. If $\tau_0 = \infty$, then (4.51) implies $M(t) < \lambda_0 < 1$ for all $t > 0$, which contradicts Lemma 4.12. 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 τ_k and the continuity of M,

$$M(\tau_k + t) \geq \lambda_k \text{ for all } t \in [0, \infty). \tag{4.52}$$

Then for a.e. $t \geq 0$, by (4.43), (4.42), (4.52), and the relations $M(t) \geq 0$ and $u^s(0) = g^s(0)$, we have

$$Z'(\tau_k + t) = M(\tau_k + t)g^s(0) + \int_0^t M(\tau_k + t - w)(u^s)'(w)dw \tag{4.53}$$

$$+ \int_t^{\tau_k + t} M(\tau_k + t - w)(u^s)'(w)dw \tag{4.54}$$

$$\leq M(\tau_k + t)g^s(0) + \lambda_k \left(u^s(t) - g^s(0)\right). \tag{4.55}$$

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 [44] we can conclude that $\lim_{t\to\infty} u^s(t) = 1$. Hence, there exists $\sigma_k > 0$ such that

(4.56)
$$(\lambda - \varepsilon) + \lambda_k(u^s(t) - 1) = (\lambda - \varepsilon) + \lambda_k \left(g^s(0) - 1\right) + \lambda_k \left(u^s(t) - g^s(0)\right) < \lambda \text{ for all } t \geq \sigma_k.$$

We now show that the following statement cannot hold:

$$M(\tau_k + t) < \lambda_{k+1} = \frac{\lambda - \varepsilon}{g^s(0)} + \lambda_k \left(1 - \frac{1}{g^s(0)}\right) \text{ for all } t > \sigma_k, \tag{4.57}$$

where the equality follows from (4.47). Indeed, if this were true, then this would imply that $M(t) < \lambda_{k+1} < 1$ for all $t \geq \sigma'_k := \tau_k + \sigma_k$, which contradicts Lemma 4.12. Thus, (4.57) does not hold or, in other words, there exists $\tau'_k \in (\sigma'_k, \infty)$ such that $M(\tau'_k) \geq \lambda_{k+1}$. We now show that for a.e. $t \in (0, \infty)$, if $M(\tau'_k + t) < \lambda_{k+1}$, then $M'(\tau'_k + t) \geq 0$. Indeed, if the first inequality is true, then substituting this into (4.55) with τ'_k in place of τ_k, and using (4.56), it follows that $Z'(\tau'_k + t) < \lambda$. When combined with (4.45) the latter implies $M'(\tau'_k + t) \geq 0$. Hence (applying Lemma B.1 with $f = M$, $c = \lambda_{k+1}$, $T = \tau'_k$, $S = \infty$), it follows that $M(\tau'_k + t) \geq \lambda_{k+1}$ 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, (4.49) holds, and if $n^* < \infty$, then also $\tau_{n^*} < \infty$ and (4.49) holds with $n = n^*$. This completes the proof of the lemma. ☐

We are now in a position to present the proof of Proposition 4.11.

Proof of Proposition 4.11. We first prove the proposition when $\lambda = 1$. For this, we consider two cases.

Case 1a: $g^s(0) \le 1$. In this case, (4.46) shows that $Z'(t) \le 1$ for a.e. $t \ge 0$, then (4.45) implies that for a.e. $t \ge 0$, $M'(t) \ge 0$. Since M is absolutely continuous by (4.39) this implies that M is nondecreasing on $[0,\infty)$ and $b := \lim_{t\to\infty} M(t)$ exists. Furthermore, (4.35) 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 $t \ge 0$, $B(T_1 + t) < 1$, and thus by (2.21) $K'(T_1 + t) = 1$. Now, recalling $B(\cdot) = \langle \mathbf{1}, \nu_\cdot \rangle$ from (4.37) and combining Lemma 2.7 and Theorem 2.6, it follows that (2.20) holds with $\psi = \mathbf{1}$, and ν_t and K_t replaced with ν_{T_1+t}, and $K_{T_1+t} - K_{T_1}$, respectively, or in other words, for each $t \ge 0$,

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

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

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

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: $g^s(0) > 1$. Let $\varepsilon \in (0, 1/2)$. In this case, by (4.47),

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

Thus, by Lemma 4.13, for each $n \ge 1$, we have $\tau_n < \infty$ and so (4.49) implies $\liminf_{t\to\infty} M(t) \ge \lambda_n$ for each $n \ge 1$. By (4.50), we obtain $\liminf_{t\to\infty} M(t) \ge 1 - \varepsilon$. Sending $\varepsilon \downarrow 0$, we obtain $\liminf_{t\to\infty} M(t) \ge 1$. Since $\limsup_{t\to\infty} M(t) \le 1$ by (4.39) it follows that in fact $\lim_{t\to\infty} M(t) = 1$. When combined with (4.35) 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: $g^s(0) \le \lambda - \varepsilon$. In this case, (4.46) shows that $Z'(t) \le \lambda - \varepsilon < \lambda$ for a.e. $t \ge 0$, and hence (4.45) implies that for a.e. $t \ge 0$, $M'(t) \ge 0$. Moreover, by (4.44), we have $M'(t) = \lambda - Z'(t) \ge \varepsilon$ if $B(t) < 1$. By the definition of M in (4.35), we obtain

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

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

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

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 $t \ge T$. 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: $g^s(0)>\lambda-\varepsilon>1$. Then $n^*<\infty$ since (4.47) shows that $\lambda_n\uparrow(\lambda-\varepsilon)>1$ as $n\to\infty$. Since by Lemma 4.13, $\tau_{n^*}<\infty$, then the continuity of W dictates that $M(\tau_{n^*})=\lambda_{n^*}$. Together with (4.55) with $k=n^*$ and the fact that M is bounded by 1 due to (4.39), this and (4.43) together imply that for a.e. $t\geq 0$,

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

By the definition of n^*, we have $\lambda_{n^*}<1\leq\lambda_{n^*+1}$. Together with the definition of λ_n in (4.47), this implies that

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

Combining the above two displays, we obtain

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

Recalling that $\lim_{t\to\infty}u^s(t)=1$ and using the expression for λ_{n^*} from (4.47), 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 (4.44), we have for a.e. $t\geq 0$,

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

By the definition of M in (4.35), we obtain for a.e. $t\geq 0$,

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

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

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

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. $t\geq T'$, 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 4.14. *For $\lambda\geq 1$, under the assumptions of Proposition* 4.11*, suppose there exists $T<\infty$ such that $B(t)=1$ for all $t\geq T$. 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 $t\geq 0$, and so by relation (4.6) in Corollary 4.4 of [28], K has the representation

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

In view of the representation for the fluid age measure in (2.20), the convergence $\nu_t \Rightarrow \nu_*$ is then a direct consequence of Lemma 6.2 of [28] 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 ν_* 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. Convergence of X in the supercritical regime. When $\lambda>1$, we now establish the large-time convergence of $X(t)$ under Assumption 2.8.

PROPOSITION 4.15. *Suppose that $\lambda>1$ and Assumptions 2.1, 2.8, and 3.1 hold. Then for any solution (X,ν,η) to the fluid equations with arrival rate λ and initial condition $(X(0),\nu_0,\eta_0)\in\mathfrak{S}$,*

$$X(t)\to x_*^\lambda, \tag{4.58}$$

with x_*^λ being the unique element of $\mathcal{X}_\lambda$ in (2.25).

Proof. Fix $\lambda>1$. It is shown in Proposition 4.7 and Lemma 4.14 that there exists $T<\infty$ such that

$$B(t)=\langle \mathbf{1},\nu_t\rangle =1 \quad \text{for all } t\geq T \tag{4.59}$$

and

$$\nu_t\Rightarrow\nu_* \text{ and } \langle h^s,\nu_t\rangle\to 1 \quad \text{as } t\to\infty. \tag{4.60}$$

The relation (2.14) shows that $X(t)=Q(t)+1$ for all $t\geq T$, and (2.22) implies that for a.e. $t\geq T$,

$$K'(t)=\lambda - Q'(t) - \int_0^{Q(t)} h^r((F^{\eta_t})^{-1}(y))dy \quad\text{and}\quad K'(t)=\langle h^s,\nu_t\rangle. \tag{4.61}$$

By using a change of variables and (2.19), we see that for each $t\geq 0$,

$$\begin{aligned}\int_0^{Q(t)} h^r((F^{\eta_t})^{-1}(y))dy &= \lambda G^r(t\wedge (F^{\eta_t})^{-1}(Q(t))) \\ &\quad + \int_{[0,H^r)} 1_{[0,(F^{\eta_t})^{-1}(Q(t))]}(x+t)\frac{g^r(x+t)}{\bar{G}^r(x)}\eta_0(dx).\end{aligned} \tag{4.62}$$

We now claim that for every $\varepsilon>0$, there exists $T^\dagger(\varepsilon)<\infty$ such that

$$F^{\eta_t}(t)>\lambda\langle\mathbf{1},\eta_*\rangle-\varepsilon \quad\text{and}\quad (F^{\eta_t})^{-1}(Q(t))<t \quad \text{for all } t\geq T^\dagger(\varepsilon). \tag{4.63}$$

To prove the claim we argue by contradiction. Suppose to the contrary that there exists a sequence of times $\{u_n\}_{n\in\mathbb{N}}$ with $u_n\to\infty$ as $n\to\infty$ such that $(F^{\eta_{u_n}})^{-1}(Q(u_n))\geq u_n$ for each $n\in\mathbb{N}$. Then by (1.7) it follows that

$$Q(u_n)\geq F^{\eta_{u_n}}(u_n) \quad \text{for all } n\in\mathbb{N}. \tag{4.64}$$

Since by (2.15) and Lemma 4.1, $Q(t) \leq \langle \mathbf{1}, \eta_t \rangle \to \lambda \langle \mathbf{1}, \eta_* \rangle$ as $t \to \infty$ and $\eta_t \Rightarrow \lambda \eta_*$ as $t \to \infty$, this implies

$$\limsup_{n\to\infty} F^{\eta_{u_n}}(u_n) \leq \limsup_{n\to\infty} Q(u_n) \leq \lambda \langle \mathbf{1}, \eta_* \rangle. \tag{4.65}$$

On the other hand, fix any $m \in (0, \infty)$. Since $\eta_t \Rightarrow \lambda\eta_*$ as $t \to \infty$, and η_* is an absolutely continuous measure, $\eta_t[0,m] \to \lambda\eta_*[0,m]$ as $t \to \infty$. Hence, it follows that

$$\liminf_{t\to\infty} F^{\eta_t}(t) \geq \liminf_{t\to\infty} F^{\eta_t}(m) = \lambda \eta_*[0,m].$$

Letting $m \to \infty$ in the above display, we have

$$\liminf_{t\to\infty} F^{\eta_t}(t) \geq \lambda \langle \mathbf{1}, \eta_* \rangle. \tag{4.66}$$

Combining (4.66) and (4.67), it follows that $F^{\eta_{u_n}}(u_n) \to \lambda\langle \mathbf{1}, \eta_* \rangle$ as $n \to \infty$. Together with (4.65) and the fact established above that $\limsup_{t\to\infty} Q(t) \leq \lambda\langle \mathbf{1}, \eta_* \rangle$, this implies $Q(u_n) \to \lambda\langle \mathbf{1}, \eta_* \rangle$ as $n \to \infty$. Next, let $M \in (0, H^r)$ be such that $G^r(M) > 1 - \frac{1}{4\lambda}$ and choose $0 < \varepsilon < \lambda(\langle \mathbf{1}, \eta_* \rangle - \eta_*[0,M])$. Then, since (as argued above) $\eta_t[0,M] \to \lambda\eta_*[0,M]$ as $t \to \infty$, there exists $T(\varepsilon) > M$ such that $\eta_t[0,M] < \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon$ for all $t \geq T(\varepsilon)$. Moreover, for every $t \geq T(\varepsilon)$, if $Q(t) > \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon$, then $\eta_t[0,M] < Q(t)$ and hence $(F^{\eta_t})^{-1}(Q(t)) > M$, and therefore $t \wedge (F^{\eta_t})^{-1}(Q(t)) > M$. Thus, by (4.62), for every $t \geq T(\varepsilon)$,

$$\begin{aligned} Q(t) > \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon \quad \Rightarrow \quad & \int_0^{Q(t)} h^r((F^{\eta_t})^{-1}(y))dy \geq \lambda G^r(t \wedge (F^{\eta_t})^{-1}(Q(t))) \\ & \geq \lambda G^r(M) > \lambda - 1/4. \end{aligned}$$

Due to the second display in (4.61), together with (4.60), without loss of generality, by choosing $T(\varepsilon)$ larger if necessary, we can assume that for all $t \geq T(\varepsilon)$, $K'(t) = \langle h^s, \nu_t \rangle > 1 - 1/4$. Together with the first display in (4.61), this implies that for all $t \geq T(\varepsilon)$,

$$Q(t) > \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon \quad \Rightarrow \quad Q'(t) < -1 + 2/4 = -1/2.$$

In turn, this implies that the time

$$\tilde{T}(\varepsilon) := \inf\{t \geq T(\varepsilon) : Q(t) \leq \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon\}$$

must be finite. Then Lemma B.1, with $f = -Q$, $T = \tilde{T}(\varepsilon)$, $S = \infty$, and $c = -\lambda\langle \mathbf{1}, \eta_* \rangle + \varepsilon$, implies that $Q(t) \leq \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon$ for all $t \geq \tilde{T}(\varepsilon)$. By (4.67), there exists $T^\dagger(\varepsilon) > \tilde{T}(\varepsilon)$ such that for all $t \geq T^\dagger(\varepsilon)$, $F^{\eta_t}(t) > \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon \geq Q(t)$. Hence, $(F^{\eta_t})^{-1}(Q(t)) < t$, which proves the stated claim (4.64).

Given the claim (4.64), by (4.62) it follows that for all $t \geq T^\dagger(\varepsilon)$,

$$\int_0^{Q(t)} h^r((F^{\eta_t})^{-1}(y))dy = \lambda G^r((F^{\eta_t})^{-1}(Q(t))), \tag{4.67}$$

and thus (4.61) implies that

$$Q'(t) = \lambda - \langle h^s, \nu_t \rangle - \lambda G^r((F^{\eta_t})^{-1}(Q(t))). \tag{4.68}$$

Since Q is continuous and (as shown above) $Q(t) \leq \lambda\langle \mathbf{1}, \eta_* \rangle - \varepsilon$ for all $t \geq T^\dagger(\varepsilon)$, Q is bounded on $[0, \infty)$.

We now show that $q_* = \lim_{t\to\infty} Q(t)$ exists. Arguing by contradiction, suppose that there exist q_1 and q_2 such that

$$\liminf_{t\to\infty} Q(t) < q_1 < q_2 < \limsup_{t\to\infty} Q(t).$$

Due to the absolute continuity of Q, this implies there must exist two sequences of times $\{t_n\}_{n\in\mathbb{N}}$, and $\{s_n\}_{n\in\mathbb{N}}$ in $[T^\dagger(\varepsilon),\infty)$ with $t_n\to\infty$, $s_n\to\infty$ as $n\to\infty$ such that

$$Q(t_n)\to q_1 \quad \text{and} \quad Q(s_n)\to q_1, \quad \text{as } n\to\infty, \tag{4.69}$$

and $Q'(t_n)\geq 0$ and $Q'(s_n)\leq 0$ for all $n\in\mathbb{N}$. When combined with (4.69) and the second display of (4.60), this implies that as $n\to\infty$,

(4.70)
$$\limsup_{n\to\infty} \lambda G^r((F^{\eta_{t_n}})^{-1}(Q(t_n)))\leq \lambda-1 \quad \text{and} \quad \liminf_{n\to\infty} \lambda G^r((F^{\eta_{s_n}})^{-1}(Q(s_n)))\geq \lambda-1.$$

Since by Lemma 4.1, $\eta_t \Rightarrow \lambda\eta_*$ as $t\to\infty$, properties of the inverse function (see Theorem 13.6.3 of [46]) imply that there exists a dense subset $\mathcal{A}\subset(0,\infty)$ such that

$$(F^{\eta_{t_n}})^{-1}(y)\to\left(F^{\lambda\eta_*}\right)^{-1}(y) \text{ and } (F^{\eta_{s_n}})^{-1}(y)\to\left(F^{\lambda\eta_*}\right)^{-1}(y) \text{ as } n\to\infty \quad \text{for all } y\in\mathcal{A}.$$

Together with the continuity of G^r, (4.70), the fact that $\left(F^{\lambda\eta_*}\right)^{-1}$ is nondecreasing, and (4.71), this implies that for each $\delta_1,\delta_2>0$ such that $q_1-\delta_1$ and $q_1+\delta_2$ lie in $\mathcal{A}$,

$$\lambda G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_1-\delta_1))\leq \limsup_{n\to\infty} \lambda G^r((F^{\eta_{t_n}})^{-1}(Q(t_n)))\leq\lambda-1 \tag{4.71}$$

and, likewise,

$$\lambda G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_1+\delta_2))\geq \liminf_{n\to\infty} \lambda G^r((F^{\eta_{t_n}})^{-1}(Q(t_n)))\geq\lambda-1.$$

Together, the last two inequalities imply

$$\lambda G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_1-\delta_1))\leq\lambda-1\leq\lambda G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_1+\delta_2)).$$

Recalling that both G^r and $(F^{\lambda\eta_*})^{-1}$ are continuous and letting $\delta_1\vee\delta_2\downarrow 0$, it follows that

$$G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_1))=\frac{\lambda-1}{\lambda}, \tag{4.72}$$

and thus q_1+1 belongs to the set $\mathcal{X}_\lambda$ from (2.25). An exactly analogous argument can be used to show that $G^r(\left(F^{\lambda\eta_*}\right)^{-1}(q_2))=(\lambda-1)/\lambda$ or, equivalently, that $q_2+1\in\mathcal{X}_\lambda$. Since $\mathcal{X}_\lambda$ is a singleton by Assumption 2.8, this implies $q_1=q_2$, which contradicts our initial assumption $q_1<q_2$. Thus, $q_*=\lim_{t\to\infty}Q(t)$ exists.

By the same argument as above, choosing δ_1,δ_2 such that $q_*-\delta_1$ and $q_*+\delta_2\in\mathcal{A}$, and then sending $\delta_1,\delta_2\to 0$, we can conclude that (4.73) also holds when q_1 is replaced with q_*. Hence, by Assumption 2.8, $q_*+1=x_*^\lambda$, the unique element of $\mathcal{X}_\lambda$. Together with (2.14), (2.18), and (4.59), this implies that $X(t)\to q_*+1=x_*^\lambda$. This completes the proof of the proposition. □

4.5. Uniqueness of the invariant distribution. We now show how the convergence results of the last three sections can be bootstrapped to conclude, under Assumption 2.8, the uniqueness of the invariant distribution.

PROPOSITION 4.16. *Suppose $\lambda \ge 1$, suppose Assumptions* 2.1, 2.8, *and* 3.1 *hold, and suppose that for any solution (X, ν, η) to the fluid equations with arrival rate λ and initial condition $(X(0), \nu_0, \eta_0) \in \mathfrak{S}$,*

$$\eta_t \Rightarrow \lambda \eta_* \quad \textit{and} \quad B(t) \to 1. \tag{4.73}$$

Then any invariant distribution μ for the fluid equations with arrival rate λ satisfies $\mu = \delta_{z_*^\lambda}$, where $z_*^\lambda = (x_*^\lambda, \nu_*, \lambda\eta_*)$, with x_*^λ being the unique element of $\mathcal{X}_\lambda$ in (2.25).

Proof. Fix $\lambda \ge 1$ and let μ be an invariant distribution for the fluid equations with arrival rate λ. 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 μ and let (X, ν, η) be the solution to the fluid equations with arrival rate λ 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 η_t and ν_t are invariant in t since μ is an invariant distribution, 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 $t \ge 0$. Then by Lemma 4.14 and Proposition 4.7, it follows that $\nu_t \Rightarrow \nu_*$ as $t \to \infty$. Since the law of ν_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 $t \ge 0$, $B(t) = 1$, and $\eta_t = \lambda\eta_*$, the relations (2.14) and (2.13) show that almost surely for all $t \ge 0$, $X(t) = Q(t) + 1$ and

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

Moreover, using the fact that almost surely for each $t \ge 0$, $\nu_t = \nu_*$, and hence, $D(t) = t\langle h^s, \nu_* \rangle = t$, we have from (2.14), (2.12) and the fact that $E = E^\lambda$ that almost surely for each $t \ge 0$,

$$Q(t) = Q(0) + (\lambda - 1)t - \lambda \int_0^t G^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(u)))du \tag{4.74}$$

$$= Q(0) + \int_0^t \left(\lambda \bar{G}^r(\left(F^{\lambda\eta_*}\right)^{-1}(Q(u))) - 1 \right) du. \tag{4.75}$$

We now consider two cases.

Case 1: $\lambda = 1$. In this case, we have

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

It is clear from (4.76) that Q is nonincreasing on $[0, \infty)$. By the nonnegativity 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 $G^r(\left(F^{\eta_*}\right)^{-1}(q_*)) = 0$ since, otherwise, $Q(t) \to -\infty$ as $t \to \infty$, which contradicts the nonnegativity of Q. Therefore, by Assumption 2.8, the definition of $\mathcal{X}_\lambda$ in (2.25), and the fact that $\lambda - 1 = 0$, it follows that $x_* = 1$ (then $q_* = 0$) is equal to the unique element x_*^λ of $\mathcal{X}_\lambda$. As before, since μ is an invariant distribution, this implies that $\mathbb{P}(X(0) = x_*^\lambda) = 1$.

Case 2: $\lambda > 1$. In this case, it follows from Proposition 4.15 that $X(t) \to q_* + 1 = x_\lambda^*$ as $t \to \infty$. Thus, we can argue as in Case 1 that $\mathbb{P}(X(0) = x_\lambda^*) = 1$. This completes the proof of the proposition. ∎

4.6. Proof of Theorem 3.2. We now indicate where each part of the result has been proved. The proof of statement (1) is given in section 4.1, and for all $\lambda \geq 0$, the (weak) convergence of η_t to $\lambda\eta_*$ under Assumption 3.1 follows from Lemma 4.1. When $\lambda > 1$, (3.1) and (3.2) follow from Proposition 4.7 and Remark 4.6 when Assumption 3.1(1) holds, and from Proposition 4.11 and Lemma 4.14 when Assumption 3.1(2) holds, and moreover, (3.3) follows from Proposition 4.15. Further, when $\lambda = 1$ the convergence $B(t) \to 1$ under Assumption 3.1(2) also follows from Proposition 4.11. Last, the uniqueness results for the invariant distribution for λ stated in (2)(b) and (3) follow from the convergence results in (2)(a) and (3) and Proposition 4.16.

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 [10] and convergence at the fluid scale, uniformly on compact time intervals, was established. Here, we study the large-time behavior under the additional assumption that the service time hazard rate satisfies Assumption 3.1. 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 [10], which is not expected to hold beyond the exponential reneging case. As shown in [10], 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 proof of the convergence result in [10] (specifically Theorem 4.3 therein) has the same gap as that described for the single-class case in Remark 3.4; namely, from the proof in [10] one can only deduce that the invariant distributions of the N-server systems exist and are tight and that any subsequential limit of the sequence of invariant distributions must be an *invariant distribution* 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 invariant distribution (which must then coincide with the Dirac delta mass at the unique element of the invariant manifold), it suffices to establish the large-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 gap in the main optimality result of [10] 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 more general. We present the fluid equations in section 5.1, and then state and prove the theorem in section 5.2.

5.1. Fluid 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 analogues 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 analogues of cumulative class i departures from service, cumulative entries to service, and cumulative reneging, and let ν_i represent the fluid analogue 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 η_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 ith 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 $\lambda_i > 0, i = 1,\dots,J$.

Before we present the fluid equations, let us comment on the special form that the single-server fluid equations (of Definition 2.3) take when the reneging is exponential. In this case, the reneging hazard rate function is constant, namely $h^r(x) = \theta$ for all $x \geq 0$, and hence, (2.13) takes the form

$$R(t) = \theta \int_0^t Q(u)du.$$

Thus, there is no longer any need to keep track of the potential reneging measure η. Accordingly, in the multiclass setting, our fluid system is an extension of a modified set of fluid equations where the equation for R is similar to the above display, and from which η 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 λ and θ if* (5.1)–(5.2) *below are satisfied: For $\varphi \in \mathcal{C}_c^1([0,\infty)\times\mathbb{R}_+)$, and $t \geq 0$,*

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_{i,t}\rangle &= \langle \varphi(\cdot,0),\nu_{i,0}\rangle + \int_0^t \langle \varphi_x(\cdot,u) + \varphi_u(\cdot,u), \nu_{i,u}\rangle du \\ &\quad - \int_0^t \langle h(\cdot)\varphi(\cdot,u), \nu_{i,u}\rangle du + \int_0^t \varphi(0,u)dK_i(u),\end{aligned} \tag{5.1}$$

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

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

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

$$B_i(t) = B_i(0) - D_i(t) + K_i(t), \tag{5.3}$$

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

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

as well as conditions imposing work conservation and nonpreemptive priority:

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

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

Under the assumption of a bounded reneging hazard rate function, which is indeed fulfilled when the reneging distribution is exponential, it was shown in Theorem 3.1 of [10] that uniqueness holds for solutions to 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 N-server 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 [28] that the measure-valued age equation (5.1) implies that for every $\psi \in \mathcal{C}_b([0,\infty))$ or $\psi = h$,

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

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 $\psi = h$,

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

In other words, (2.20) holds with (ν, K) and G^s replaced with $(\tilde{\nu}, \tilde{K})$ and G. We now argue that $\tilde{K}$ and $\tilde{B}$ satisfy the analogue of (2.21). 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 $\tilde{K}'(u) = \tilde{\lambda}$ holds at each u in the interval. Combining this with

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

which follows from (5.3) and (5.2), we obtain, exactly as in Theorem 3.2 of [10], that for almost every 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 ρ_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, ν) to the multiclass fluid equations with initial condition (X_0, ν_0) and arrival and reneging rate vectors λ and θ 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 [10] 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 η 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 4.7 and Lemma 4.14, $\langle h, \tilde{\nu}_t \rangle \to 1$. As a result, by (2.21), 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 $i_0 \in \{1, \dots, J-1\}$ 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

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

(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

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

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 ℓ, $\hat{\lambda} = \sum_{i=1}^{\ell} \lambda_i < 1$, 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 nonempty. Since $\hat{Q}$ is continuous, $\mathcal{O}$ is open and is a union of countable open intervals. For a.e. u in each such interval, by (5.7), for all $i > \ell$, $K_i'(u) = 0$. 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 $\tilde{B}(u) = 1$. In turn, (5.3), (5.5), (5.2), and the fact that $\tilde{R}$ is nondecreasing together imply that for a.e. $u \in \mathcal{O}$, $\tilde{D}'(u) = \tilde{K}'(u) = \langle h, \tilde{\nu}_u \rangle$. Thus, we have for a.e. $u > t_0$,

$$\text{if} \quad \hat{Q}(u) > 0 \quad \text{then} \quad \hat{Q}'(u) = \hat{\lambda} - \hat{R}'(u) - \hat{K}'(u) \leq \hat{\lambda} - \langle h, \tilde{\nu}_u \rangle \leq -\varepsilon_0.$$

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 $\nu_{i,t}(dx)$ 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 ρ_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 $\mathcal{O}_m := \{u \in [t_3, \infty) : Q_m(u) > 0\}$. We show below that $\mathcal{O}_m = [t_3, \infty)$. Toward 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

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

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

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

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 $i > m$. We can therefore use a version of (2.20) for ν_i to conclude that $\nu_{i,t} \Rightarrow 0$ and $B_i(t) \to 0$ for $i > m$. 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 $[0,\infty)$,

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

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

Step 3. Last, 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 $i \le m$, 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 ψ 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 $\nu_i(t)$, $i \le \ell = m-1$, 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 $i \le \ell$, (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. Recall $z_f := \int_0^\infty h^s(x) f(x) dx < \infty$ and let $U(x) := x \log x$, $x > 0$, $U(0) = 0$. Fix a measurable function $f : [0,\infty) \mapsto \mathbb{R}_+$ with $\int_0^\infty f(x)dx \le 1$. For notational conciseness, define

$$\begin{aligned} A(f) &:= \int_0^\infty h^s(x) f(x) \log \frac{f(x)}{f_*(x)} dx - z_f \log z_f \\ &= \int_0^\infty U\left(\frac{f(x)}{f_*(x)}\right) h^s(x) f_*(x) dx - U(z_f). \end{aligned} \tag{A.1}$$

Since $\int_0^\infty h^s(x) f_*(x) dx = 1$ and $\int_0^\infty \frac{f(x)}{f_*(x)} h^s(x) f_*(x) dx = z_f$, the convexity of U and Jensen's inequality imply the nonnegativity of $A(f)$. 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

$$\begin{aligned} A(f) &= \int_0^\infty V\Big(\frac{f(x)}{f_*(x)}\Big) h^s(x) f_*(x)dx + \int_0^\infty U'(z_f)\Big(\frac{f(x)}{f_*(x)} - z_f\Big) h^s(x) f_*(x)dx \\ &= \int_0^\infty V\Big(\frac{f(x)}{f_*(x)}\Big) h^s(x) f_*(x)dx, \end{aligned}$$

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

$$\begin{aligned} A(f) &\geq \varepsilon^s \int_0^\infty V\Big(\frac{f(x)}{f_*(x)}\Big) f_*(x)dx \\ &= \varepsilon^s \left[\int_0^\infty f(x)\log\frac{f(x)}{f_*(x)}dx - \int_0^\infty U'(z_f)\Big(\frac{f(x)}{f_*(x)} - z_f\Big) f_*(x)dx - U(z_f)\right] \\ &= \varepsilon^s [R(\mu^f \| \nu_*) - c_f U'(z_f) + U'(z_f) z_f - U(z_f)] \\ &= \varepsilon^s [R(\mu^f \| \nu_*) - c_f \log z_f - c_f + z_f] \\ &= \varepsilon^s \{R(\mu^f \| \nu_*) + c_f[-\log z_f - 1 + z_f] + z_f(1 - c_f)\} \\ &\geq \varepsilon^s R(\mu^f \| \nu_*), \end{aligned}$$

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 almost every $t \in [T, \infty)$, $f'(t) \geq 0$ if $f(t) < c$. Hence the lemma is proved. □

Acknowledgments. We would like to thank Amber Puha for raising the question of a full justification of the convergence results in Theorem 3.3 of [27]. We would also like to thank the referees for their careful reading of the paper and for useful suggestions that helped improve and streamline the exposition.

REFERENCES

[1] P. Agarwal and K. Ramanan, *Invariant States of Hydrodynamic Limits of Randomized Load Balancing Networks*, preprint, https://arxiv.org/abs/2008.08510, 2020.

[2] R. Aghajani and K. Ramanan, *Hydrodynamic limits of randomized load balancing networks*, Ann. Appl. Probab., 29 (2019), pp. 2114–2174.

[3] R. Aghajani and K. Ramanan, *Ergodicity of an SPDE associated with a many-server queue*, Ann. Appl. Probab., 29 (2019), pp. 994–1045.

[4] R. Aghajani and K. Ramanan, *The limit of stationary distributions of many-server queues in the Halfin-Whitt regime*, Math. Oper. Res., 45 (2020), pp. 1016–1055.

[5] R. Aghajani, X. Li, and K. Ramanan, *The PDE method for the analysis of randomized load balancing networks*, Proc. ACM Meas. Anal. Comput. Syst., 1 (2017), 38.

[6] B. E. Ainseba, S. Anita, and M. Langlais, *Optimal Control for a Nonlinear Age-Structured Population Dynamics Model*, Southwest Texas State University, 2003.

[7] S. Asmussen, *Applied Probability and Queues*, 2nd ed., Springer, New York, 2003.

[8] R. Atar, C. Giat, and N. Shimkin, *The $c\mu/\theta$ rule for many-server queues with abandonment*, Oper. Res., 58 (2010), pp. 1427–1439.

[9] R. Atar, C. Giat, and N. Shimkin, *On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost*, Queueing Syst., 67 (2011), pp. 127–144.

[10] R. Atar, H. Kaspi, and N. Shimkin, *Fluid limits for many-server systems with reneging under a priority policy*, Math. Oper. Res., 39 (2013), pp. 672–696.

[11] F. Baccelli and G. Hebuterne, *On queues with impatient customers*, in Performance' 81, E. Gelenbe, ed., North-Holland, Amsterdam, 1981, pp. 159–179.

[12] M. B. Brown, *Inequalities, and monotonicity properties for some specialized renewal processes*, Ann. Probab., 8 (1980), pp. 227–240.

[13] J. A. Cañizo, J. A. Carrillo, and S. Cuadrado, *Measure solutions for some models in population dynamics*, Acta Appl. Math., 123 (2013), pp. 141–156.

[14] E. Cinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

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

[16] J. M. Cushing, *An Introduction to Structured Population Dynamics*, SIAM, Philadelphia, 1998.

[17] J. G. Dai, A. B. Dieker, and X. Gao, *Validity of heavy-traffic steady-state approximations in many-server queues with abandonment*, Queueing Syst., 78 (2014), pp. 1–29.

[18] J. G. Dai and S. He, *Many-server queues with abandonment: A survey of diffusion and fluid approximations*, J. Syst. Sci. Syst. Eng., 21 (2012), pp. 1–36.

[19] A. Ducrot, P. Magal, and O. Seydi, *Nonlinear boundary conditions derived by singular perturbation in age structured population dynamics model*, J. Appl. Anal. Comput., 1 (2011), pp. 373–395.

[20] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley & Sons, New York, 2011.

[21] N. Fournier and B. Perthame, *A nonexpanding transport distance for some structured equations*, SIAM J. Math. Anal., 53 (2021), pp. 6847–6872.

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

[23] D. Gamarnik and D. A. Goldberg, *Steady-state $GI/G/n$ queue in the Halfin-Whitt regime*, Ann. Appl. Probab., 23 (2013), pp. 2382–2419.

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

[25] O. Garnett, A. Mandelbaum, and M. Reiman, *Designing a call center with impatient customers*, Manuf. Serv. Oper. Manag., 4 (2002), pp. 208–227.

[26] W. Kang and K. Ramanan, *Fluid limits of many-server queues with reneging*, Ann. Appl. Probab., 20 (2010), pp. 2204–2260.

[27] W. Kang and K. Ramanan, *Asymptotic approximations for stationary distributions of many-server queues with abandonment*, Ann. Appl. Probab., 22 (2012), pp. 477–521.

[28] H. Kaspi and K. Ramanan, *Law of large numbers limits for many-server queues*, Ann. Appl. Probab., 21 (2011), pp. 33–114.

[29] Z. Long and J. Zhang, *Virtual allocation policies for many-server queues with abandonment*, Math. Methods Oper. Res., 90 (2019), pp. 399–451.

[30] Z. Long, N. Shimkin, H. Zhang, and J. Zhang, *Dynamic scheduling of multiclass many-server queues with abandonment: The generalized $c\mu/\theta$ rule*, Oper. Res., 68 (2020).

[31] A. Mandelbaum and S. Zeltyn, *Staffing many-server queues with impatient customers: Constraint satisfaction in call centers*, Oper. Res., 57 (2009), pp. 1189–1205.

[32] W. H. Mather, N. Cookson, J. Hasty, L. S. Tsimring, and R. J. Williams, *Correlation resonance generated by coupled enzymatic processing*, Biophys. J., 99 (2010), pp. 3172–3181.

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

[34] S. Mischler, B. Perthame, and L. Ryzhik, *Stability in a nonlinear population maturation model*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1751–1772.

[35] J. A. Mulvany, A. L. Puha, and R. J. Williams, *Asymptotic behavior of a critical fluid model for a multiclass processor sharing queue via relative entropy*, Queueing Syst., 93 (2019), pp. 351–397.

[36] O. Nakoulima, A. Omrane, and J. Velin, *A nonlinear problem for age-structured population dynamics with spatial diffusion*, Topol. Methods Nonlinear Anal., 17 (2001), pp. 307–319.

[37] C. P. Niculescu and L.-E. Persson, *Convex Functions and Their Applications: A Contemporary Approach*, Springer-Verlag, New York, 2005.

[38] F. Paganini, A. Tang, A. Ferragut, and L. L. H. Andrew, *Network stability under alpha fair bandwidth allocation with general file size distribution*, IEEE Trans. Automat. Control, 57 (2012), pp. 579–591.

[39] B. Perthame, *Transport Equations in Biology*, Front. Math., Birkhäuser, Basel, 2007.

[40] B. Perthame, *Introduction to Structured Equations in Biology*, CNA Summer School Lecture Notes, 2008.

[41] B. Perthame and S. K. Tumuluri, *Nonlinear renewal equations*, in Selected Topics in Cancer Modeling, Birkhäuser, Boston, 2008, pp. 1–32.

[42] A. L. Puha and A. Ward, *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, pp. 189–217.

[43] A. L. Puha and R. J. Williams, *Asymptotic behavior of a critical fluid model for a processor sharing queue via relative entropy*, Stoch. Syst., 6 (2016), pp. 251–300.

[44] W. L. Smith, *Asymptotic renewal theorems*, Proc. Roy. Soc. Edinburgh Sect. A, 64 (1954), pp. 9–48.

[45] S. Zeltyn and A. Mandelbaum, *Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue*, Queueing Syst., 51 (2005), pp. 361–402.

[46] W. Whitt, *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*, Springer, New York, 2002.