MATHEMATICS OF OPERATIONS RESEARCH
Vol. 33, No. 2, May 2008, pp. 375–402
ISSN 0364-765X | EISSN 1526-5471 | 08 | 3302 | 0375

informs®

DOI 10.1287/moor.1070.0298

Fluid Limits for Processor-Sharing Queues with Impatience

H. Christian Gromoll
Department of Mathematics, University of Virginia, Charlottesville, Virginia 22903, USA,
gromoll@virginia.edu, http://www.faculty.virginia.edu/gromoll/

Philippe Robert
INRIA-Rocquencourt, RAP Project, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France,
Philippe.Robert@inria.fr, http://www-rocq.inria.fr/~robert

Bert Zwart
Georgia Institute of Technology, Stewart School of ISyE, 765 Ferst Drive, Atlanta, Georgia 30332, USA,
bertzwart@gatech.edu

We investigate a processor-sharing queue with renewal arrivals and generally distributed service times. Impatient jobs may abandon the queue or renege before completing service. The random time representing a job's patience has a general distribution and may be dependent on its initial service time requirement. A scaling procedure that gives rise to a fluid model with nontrivial yet tractable steady state behavior is presented. This fluid model captures many essential features of the underlying stochastic model, and it is used to analyze the impact of impatience in processor-sharing queues.

Key words: processor sharing; queues with impatience; measure-valued process; fluid limits; delay-differential equations; empirical processes
MSC2000 subject classification: Primary: 60K25, 60K30, 60G57, 60F17; secondary: 90B15, 90B22
OR/MS subject classification: Primary: queues, limit theorems; secondary: scheduling
History: Received April 5, 2006; revised April 30, 2007, and August 30, 2007.

1. Introduction.

Processor-sharing policy and impatience. Processor-sharing (PS) policies were originally proposed as models of time sharing in computer operating systems. Recently, generalizations of this discipline have been used to describe data transfers in congested routes through the Internet (see Roberts and Massoulié [25] and Kelly and Williams [18] and references therein). This has created considerable renewed interest in the analysis of PS policies.

This paper studies the behavior of a $GI/GI/1$ queue serving impatient jobs according to the PS policy: If there are N jobs in the queue, each job receives simultaneous service at rate $1/N$. An *impatient job* has a random *initial lead time* in addition to its service time. Such a job has a *deadline* equal to its arrival time plus its initial lead time; if the job has not completed service when the deadline expires, it abandons the queue (or *reneges*) and therefore does not complete service. For example, the timeout of a transmission control protocol (TCP) flow through the Internet can be thought of as the expiration of a random deadline and subsequent reneging of the flow.

The impact of impatience on PS queues is greater than for first-in-first-out (FIFO) queues. A typical job that abandons a FIFO queue will do so while waiting to begin service. In contrast, a job that abandons a PS queue will have already received partial service. Because this partial service is wasted, impatience may create significant overhead for a PS server.

There is a large literature on queueing models with impatience under the FIFO discipline. An early paper by Barrer [1] considers an example arising in a military application. Stanford [27] surveys the literature in this domain (see also Stanford [26] and Boots and Tijms [5]). This body of work focuses primarily on exact performance analysis. Ward and Glynn [29] have recently obtained a diffusion approximation for single channel queues. There are also various studies of multiserver queues with abandonments, motivated by call center applications; see the survey by Bonald and Massoulié [3], Gans et al. [10], and references therein.

There is some related literature treating other policies, but in the context of *soft deadlines*. Jobs with soft deadlines are not impatient; they remain in the system until completing service, even if their deadlines have expired. In particular, these queues are work conserving in the sense that the server must fully process all work arriving to the system. Results for such models describe the extent to which overdue jobs are produced by the underlying service discipline, without the effect of abandonments. Doytchinov et al. [9] and Kruk et al. [20, 21] investigate the heavy traffic behavior of various systems using the earliest deadline first and FIFO policies. Gromoll and Kruk [12] describe the heavy traffic behavior of a PS queue incorporating a fairly general structure of soft deadlines.

For PS queues with impatience, only a few results are known. Coffman et al. [7] cover the special case of exponential service times and lead times, where the lead time and service time are independent. Guillemin et al. [14] consider heavy tailed service times, and obtain some results on the reneging behavior of large jobs by analyzing the tail behavior of the sojourn time distribution. Using some approximations, Bonald and Roberts [4] analyze the steady state of a system with general service times and some dependence between service times and lead times.

Results of the paper. This paper analyzes the PS queue with impatience using fluid limits. The dynamics of the system are represented as a measure-valued process: the system state at time $t \geq 0$ is represented as a random point measure $\mathcal{Z}(t)$ on $(0, \infty] \times (0, \infty]$, such that $\mathcal{Z}(t)$ has a point mass at $(b, d) \in (0, \infty] \times (0, \infty]$ if and only if there is a job in the system at time t with residual service time b and residual lead time d. See Jean-Marie and Robert [16] and Doytchinov et al. [9] for similar representations of residual service times in single server queues. This setup enables a fairly general analysis. The case of a general joint distribution of service times and initial lead times with possible dependence of the two random variables is included in our setting.

Under mild assumptions, we show that under a convenient scaling, a family of measure-valued processes associated with $(\mathcal{Z}(t))$ is tight and converges in distribution to a limit $(\zeta(t))$. For $t \geq 0$, $\zeta(t)$ is a nonnegative measure on $(0, \infty] \times (0, \infty]$. This fluid limit is characterized as the solution of a functional Equation (2.9), which can be viewed as a time-changed functional differential equation.

The overloaded case $\rho > 1$, which is our main focus, presents a nontrivial and interesting steady state behavior. The total fluid mass in the system at equilibrium (the fluid analogue of the total number of jobs) is shown to be the solution z_∞ of a simple fixed point Equation (3.4). Moreover, the steady state of the fluid model, that is, the limit of $\zeta(t)$ as t goes to infinity, is a measure on $(0, \infty] \times (0, \infty]$, which has a simple expression (2.15) in terms of z_∞.

These results provide significant insight into the qualitative properties of PS queues with impatience. An interpretation of the fixed point Equation (3.4) is given and used to analyze the total number of jobs in the system, and to estimate the long run fraction of jobs that renege. The impact of the variability of service times and lead times, as well as other properties of this model, are investigated in Gromoll et al. [11].

In contrast to previous work on PS models, the server considered here might only process a fraction of the service requirement of a job. This creates an important difference: The workload process does not coincide with that of a FIFO queue, a fact that previous work has exploited heavily. For this reason, analysis of the fluid model is more intricate. A different approach to prove existence, uniqueness, and convergence to steady state of fluid model solutions is used. It is shown that there exists a *maximal* fluid model solution and, using monotonicity arguments, the properties of fluid limits are investigated under fairly general assumptions.

Organization. A detailed description of the model and main results are presented in §2. Qualitative properties of the fluid model are analyzed in §3. Section 4 is devoted to examples. Sections 5 and 6 are concerned with convergence to the fluid limit. Section 5 establishes tightness, and §6 characterizes limit points.

2. Model description and results.

This section gives a detailed description of the stochastic processes associated to this queue, as well as a summary of our main results.

2.1. Stochastic model.

The stochastic model consists of the following: a processor-sharing server working at unit rate from an infinite capacity buffer, a collection of stochastic primitives $E(\cdot)$, $\{B_i, D_i\}$ describing, respectively, the arrival process, service requirements, and initial lead times of jobs, and a random initial condition specifying the state of the system at time 0. All random objects are defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with expectation operator $\mathbf{E}[\cdot]$.

The *exogenous arrival process* $(E(t), t \geq 0)$ has rate $\lambda > 0$; it is a delayed renewal process starting from zero, with ith jump time U_i. For $t \geq 0$, $E(t)$ is the number of jobs that arrive to the buffer during $(0, t]$. For $i \geq 1$, U_i is the arrival time of job i and $\lambda = 1/\mathbf{E}[U_2]$ is the arrival rate. Jobs already in the buffer at time 0 are called *initial jobs*.

For $i \geq 1$, the *service time* B_i is a strictly positive random variable representing the amount of processing time that job i requires from the server. The random variable D_i is strictly positive and determines the deadline of job i: It represents the maximum amount of time that job i will stay in the buffer. Because job i arrives at time U_i, its deadline is at time $U_i + D_i$. It will abandon the system at this time if it has not yet completed service. The random variable D_i is called the *initial lead time* of job i.

The model allows either the service time or the initial lead time (but not both) to be equal to infinity. In this way, the ordinary PS queue ($D_i \equiv \infty$) and the infinite server queue ($B_i \equiv \infty$) are included as special cases. Therefore, the random vector (B_i, D_i) takes values in the space $\overline{\mathbb{R}}_+^2 = [0, \infty] \times [0, \infty]$. Here, $\overline{\mathbb{R}}_+ = [0, \infty]$ is the disjoint union $[0, \infty) \cup \{\infty\}$, with the arithmetic extensions $x + \infty = \infty$ for all $x \in \overline{\mathbb{R}}_+$, $x \cdot \infty = \infty$ for $x > 0$, and $0 \cdot \infty = 0$. (Note that the space $\overline{\mathbb{R}}_+ = [0, \infty]$ as used in this paper is distinct from the one-point compactification of $[0, \infty)$.) The collection of Borel subsets of $\overline{\mathbb{R}}_+^2$ is denoted by $\mathscr{B}$. Throughout the paper, it is assumed that all sequences of service times and initial lead times $\{B_i, D_i\}$ are independent and identically distributed (i.i.d.) $\overline{\mathbb{R}}_+^2$-valued random variables, and that their common joint distribution ϑ on $\overline{\mathbb{R}}_+^2$ satisfies

$$\vartheta(\{0\} \times \overline{\mathbb{R}}_+) = \vartheta(\overline{\mathbb{R}}_+ \times \{0\}) = \vartheta((\infty, \infty)) = 0. \tag{2.1}$$

Note that the random variables B_i and D_i may be dependent. A generic random element of $\overline{\mathbb{R}}_+^2$ with distribution ϑ will be denoted (B, D).

Initial condition. The *initial condition* specifies $Z(0)$, the number of initial jobs present in the buffer at time zero, as well as the service times and initial lead times of these initial jobs. Assume that $Z(0)$ is a nonnegative, integer-valued random variable. The service times and initial lead times for initial jobs are the first $Z(0)$ elements of a sequence $\{B_j^0, D_j^0\}$ of random variables taking values in $(0, \infty] \times (0, \infty] \backslash (\infty, \infty)$ almost surely. Assume that the expected number of initial jobs is finite: $\mathbf{E}[Z(0)] < \infty$.

Time evolution of the queue. For each $t \geq 0$, let $Z(t)$ denote the number of jobs in the buffer (or *queue length*) at time t, and let $S(t)$ denote the *cumulative service per job* provided by the server up to time t. Because of the processor sharing policy, $S(t)$ is given by

$$S(t) = \int_0^t \frac{1}{Z(s)} \, \mathrm{d}s, \tag{2.2}$$

where the integrand is defined to be zero when the queue length equals zero. If a job arrived at time $s \geq 0$ and is still present in the queue at time $t \geq s$, then by time t it has received the cumulative amount of processing time $S(s, t) = S(t) - S(s)$.

Therefore, the *residual service time* at time t of job $i \leq E(t)$ and initial job $j \leq Z(0)$ are given by

$$B_i(t) = (B_i - (S(t) - S(U_i)))^+ \qquad \text{and} \qquad B_j^0(t) = (B_j^0 - S(t))^+. \tag{2.3}$$

Define the *lead time* at time t of job $i \leq E(t)$ and initial job $j \leq Z(0)$ by

$$D_i(t) = (U_i + D_i - t)^+ \qquad \text{and} \qquad D_j^0(t) = (D_j^0 - t)^+. \tag{2.4}$$

A job's residual service time is the remaining amount of processing time required to fulfill its service requirement; its lead time is the remaining time until its deadline. Job i will depart the system either when its service requirement is fulfilled or when its deadline is reached; it will leave the system at time

$$\inf\{t \geq U_i\colon\ \min\{B_i(t), D_i(t)\} = 0\}.$$

The *state descriptor* is a measure-valued process that keeps track of the residual service times and lead times of all jobs in the buffer. For job i, this information is represented as a unit of mass at the point $(B_i(t), D_i(t)) \in \overline{\mathbb{R}}_+^2$ at all times $t \geq U_i$ such that job i is still in the system. Let $\delta_{(x,y)}^+$ denote the Dirac point measure at $(x, y) \in \overline{\mathbb{R}}_+^2$ if $\min\{x, y\} > 0$, otherwise $\delta_{(x,y)}^+$ is the zero measure. Then, the state of the system at time $t \geq 0$ is represented by the random point measure

$$\mathscr{Z}(t) = \sum_{j=1}^{Z(0)} \delta^+_{(B_j^0(t),\, D_j^0(t))} + \sum_{i=1}^{E(t)} \delta^+_{(B_i(t),\, D_i(t))}. \tag{2.5}$$

Note that the queue length at time t is given by the total mass of the measure $\mathscr{Z}(t)$,

$$Z(t) = \langle 1, \mathscr{Z}(t) \rangle, \tag{2.6}$$

where $\langle f, \mu \rangle = \int_{\overline{\mathbb{R}}_+^2} f d\mu$ for a Borel measure μ on $\overline{\mathbb{R}}_+^2$ and a μ-integrable function $f\colon \overline{\mathbb{R}}_+^2 \to \mathbb{R}$.

In this way, the dynamics of the system are represented as a distribution of point masses on $\overline{\mathbb{R}}_+^2$ moving toward the axes. At time $t \geq 0$, points move left at rate $1/Z(t)$ and down at rate 1. (A point with one coordinate

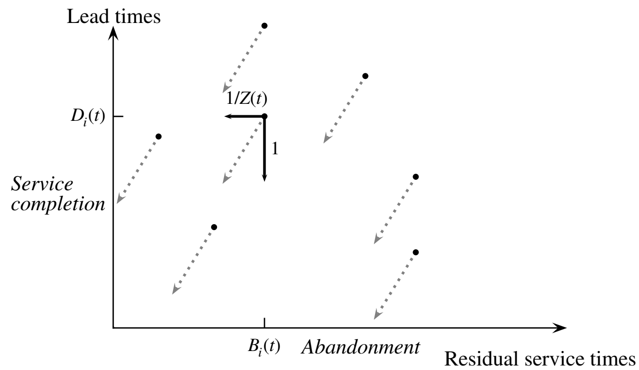

FIGURE 1. Dynamics of the measure-valued process $\mathcal{Z}(\cdot)$.

equal to infinity will remain that way while the other coordinate moves.) Point masses vanish when hitting one of the axes: a point mass reaching the vertical axis corresponds to a job completing service, while a point mass hitting the horizontal axis represents a job abandoning the queue (see Figure 1).

Let $\mathbf{M}$ denote the space of finite nonnegative Borel measures on $\overline{\mathbb{R}}_+^2$, endowed with the topology of weak convergence: $\zeta_n \xrightarrow{\mathbf{w}} \zeta$ in $\mathbf{M}$ if and only if $\langle f, \zeta_n\rangle \to \langle f, \zeta\rangle$ for all continuous bounded functions $f\colon \overline{\mathbb{R}}_+^2 \to \mathbb{R}$. Let $\mathbf{D}([0,\infty), \mathbf{M})$ denote the space of *càdlàg* paths in $\mathbf{M}$, endowed with the Skorohod J_1-topology. Then, for $t \geq 0$, $\mathcal{Z}(t)$ is a random element of $\mathbf{M}$ and $\mathcal{Z}(\cdot)$ is a random element of $\mathbf{D}([0,\infty), \mathbf{M})$.

It is clear that, given stochastic primitives $E(\cdot)$, $\{B_i, D_i\}$, and the initial condition $\mathcal{Z}(0)$, the Equations (2.2)–(2.6) uniquely determine the processes $S(\cdot)$, $Z(\cdot)$, $\mathcal{Z}(\cdot)$, and the residual service times and lead times. It is also easily seen that the state descriptor $\mathcal{Z}(\cdot)$ satisfies the following equation: For each Borel set $A \in \mathcal{B}$ and all $t \geq 0$,

$$\mathcal{Z}(t)(A) = \mathcal{Z}(0)(A + (S(t), t)) + \sum_{i=1}^{E(t)} 1_A^+(B_i(t), D_i(t)), \tag{2.7}$$

where $A + w = \{a + w\colon a \in A\}$ and $1_A^+(w) = \langle 1_A, \delta_w^+\rangle$. Note that the quantity $\mathcal{Z}(0)(A + (S(t), t))$ corresponds to a shift of the initial points by the vector $(S(t), t)$: if $(x, y) \in \overline{\mathbb{R}}_+^2$ and $(s, t) \in \overline{\mathbb{R}}_+^2$, then for $A \in \mathcal{B}$,

$$\delta_{(x,y)}(A + (s, t)) = \delta_{(x-s,\, y-t)}(A).$$

Equation (2.7) plays a crucial role in determining fluid limits for the model.

2.2. A fluid scaling. We now introduce a sequence of renormalized stochastic processes $\bar{\mathcal{Z}}^r(\cdot)$ associated to the solution of the evolution Equation (2.7). The limits in distribution of $\bar{\mathcal{Z}}^r(\cdot)$ will give the fluid limits of this model.

Let $\mathcal{R} \subset [0, \infty)$ be a sequence increasing to infinity. Suppose that for each $r \in \mathcal{R}$, there is a stochastic model as defined in §2.1. That is, for each $r \in \mathcal{R}$, there are stochastic primitives $E^r(\cdot)$ and $\{B_i^r, D_i^r\}$, with associated data λ^r and ϑ^r and an initial condition $\mathcal{Z}^r(0)$. As before, these determine stochastic processes $Z^r(\cdot)$, $S^r(\cdot)$, $\mathcal{Z}^r(\cdot)$, and residual service times and lead times $\{B_i^r(\cdot), D_i^r(\cdot)\}$ and $\{B_i^{0r}(\cdot), D_i^{0r}(\cdot)\}$.

A fluid scaling is applied to each model in the sequence. To obtain nontrivial scaling limits, initial lead times $\{D_i^r\}$ will be assumed to be of order r. Consequently, they must be scaled by r^{-1} to keep track of them. For each $r \in \mathcal{R}$, let $\check{\vartheta}^r \in \mathbf{M}$ be the probability measure defined by

$$\check{\vartheta}^r(F \times G) = \vartheta^r(F \times rG)$$

for all Borel subsets F, G of $\overline{\mathbb{R}}_+$, with the notation $rG = \{r \cdot g\colon g \in G\}$. Note that if $\{B_i^r, D_i^r\} = \{B_i, rD_i\}$ for some sequence $\{B_i, D_i\}$, then $\check{\vartheta}^r$ is simply the distribution of (B_1, D_1).

For each $r \in \mathcal{R}$, the fluid scaled state descriptor is defined, for $t \geq 0$, as the random measure $\bar{\mathcal{Z}}^r(t) \in \mathbf{M}$ such that

$$\bar{\mathcal{Z}}^r(t)(F \times G) = \frac{1}{r}\mathcal{Z}^r(rt)(F \times rG)$$

for all Borel subsets F, G of $\overline{\mathbb{R}}_+$. Note that this definition scales lead times by a factor r^{-1} as well. Fluid scaled versions of the remaining processes are defined as follows: for all $r \in \mathcal{R}$, $t \geq s \geq 0$ and $i = 1, \ldots, E^r(rt)$, let

$$\bar{E}^r(t) = \frac{1}{r}E^r(rt), \qquad \bar{Z}^r(t) = \frac{1}{r}Z^r(rt),$$

$$\bar{S}^r(t) = S^r(rt), \qquad \bar{S}^r(s,t) = \bar{S}^r(t) - \bar{S}^r(s),$$
$$\bar{B}_i^r(t) = B_i^r(rt), \qquad \bar{D}_i^r(t) = \frac{1}{r} D_i^r(rt).$$

2.3. Fluid model. We next introduce a deterministic fluid model satisfying dynamic equations analogous to (2.7). It will be shown that these equations can be obtained as limits of (2.7) under the above scaling.

Let $\lambda > 0$, let $\vartheta \in \mathbf{M}$ be a probability measure satisfying (2.1), and let $\zeta_0 \in \mathbf{M}$ be such that $\zeta_0(\{\infty\}\times\{\infty\}) = 0$ and such that the projections $\zeta_0(\cdot \times \overline{\mathbb{R}}_+)$ and $\zeta_0(\overline{\mathbb{R}}_+ \times \cdot)$ are free of atoms in $[0,\infty)$. The fluid model is defined from the data $(\lambda, \vartheta, \zeta_0)$, which will arise as limiting values of the parameters $(\lambda^r, \breve{\vartheta}^r, \bar{\mathscr{Z}}^r(0))$ of the stochastic model (see Assumptions $\boldsymbol{\mathcal{A}}$ below). A random vector in $\overline{\mathbb{R}}_+$ with distribution ϑ is denoted (B, D). Let $\rho = \lambda \mathbf{E}[B]$ denote the *traffic intensity* of the fluid model. It is assumed throughout that $\rho > 1$, that is, the server is nominally overloaded.

Definition 2.1. A continuous function $\zeta(\cdot)$: $[0,\infty) \to \mathbf{M}$ is called a measure-valued fluid model solution for the data $(\lambda, \vartheta, \zeta_0)$ if

(i) $\inf_{t>a} z(t) > 0$ for all $a > 0$,

(ii) for all $A \in \mathscr{B}$ and $t \geq 0$,

$$\zeta(t)(A) = \zeta_0(A + (S(0,t), t)) + \lambda \int_0^t \vartheta(A + (S(s,t), t-s))\,\mathrm{d}s, \tag{2.8}$$

where for all $v \geq u \geq 0$,

$$S(u,v) = \int_u^v \frac{1}{z(s)}\,\mathrm{d}s,$$

and $z(\cdot)$ is the total mass function

$$z(t) = \langle 1, \zeta(t)\rangle = \int_{\overline{\mathbb{R}}_+^2} \zeta(t)(\mathrm{d}x, \mathrm{d}y).$$

The function $z(\cdot)$ is called simply a fluid model solution for $(\lambda, \vartheta, \zeta_0)$.

Note that $S(0,t)$ may be equal to ∞ if $z(0) = 0$, and thus $\zeta_0 = 0$. Both right-hand terms in (2.8) are still well-defined in this case and the first term equals zero.

Define a class of corner sets

$$\mathscr{C} = \{[x,\infty] \times [y,\infty]\colon x, y \in \overline{\mathbb{R}}_+\}.$$

The sets $C \in \mathscr{C}$ are useful for describing the evolution of fluid model solutions. Because each C of the form $[x,\infty]\times[y,\infty]$ is characterized by the coordinates (x,y) of its corner, it will be convenient to use the notation $\mu(x,y) \stackrel{\text{def.}}{=} \mu([x,\infty]\times[y,\infty])$ for any $\mu \in \mathbf{M}$. Then, (2.8) can be rewritten for this class of subsets as follows. Let $z_0 = \langle 1, \zeta_0\rangle$. Let (B_0, D_0) be a random vector with distribution ζ_0/z_0 if $z_0 > 0$, and let (B_0, D_0) be the zero vector if $z_0 = 0$. Recall that (B, D) is a random vector with distribution ϑ. Then, for each $x, y \in \overline{\mathbb{R}}_+$, and $t \geq 0$, a measure-valued fluid model solution satisfies

$$\zeta(t)(x,y) = z_0 \mathbf{P}(B^0 \geq x + S(0,t);\, D^0 \geq y + t) + \lambda \int_0^t \mathbf{P}(B \geq x + S(s,t);\, D \geq y + t - s)\,\mathrm{d}s. \tag{2.9}$$

Because $z(t) = \zeta(t)(0,0)$ for each $t \geq 0$, a fluid model solution $z(\cdot)$ satisfies

$$z(t) = z_0 \mathbf{P}(B^0 \geq S(0,t);\, D^0 \geq t) + \lambda \int_0^t \mathbf{P}(B \geq S(s,t);\, D \geq t - s)\,\mathrm{d}s. \tag{2.10}$$

Conversely, if $\zeta(\cdot)$: $[0,\infty) \to \mathbf{M}$ is a continuous function satisfying (i) of Definition 2.1 and (2.9) for all $x, y \in \overline{\mathbb{R}}_+$ and $t \geq 0$, then $\zeta(\cdot)$ is a measure-valued fluid model solution as the following argument shows.

Let $\mathscr{C}'$ be the set of $A \in \mathscr{B}$ for which (2.8) holds for all $t \geq 0$. Clearly, $\mathscr{C} \subset \mathscr{C}'$ because (2.9) holds for all $x, y \in \overline{\mathbb{R}}_+$ and $t \geq 0$. Observe that $\mathscr{C}'$ is a λ-system: $\overline{\mathbb{R}}_+^2 \in \mathscr{C}'$ because $\overline{\mathbb{R}}_+^2 \in \mathscr{C}$; if $\{A_n\} \subset \mathscr{C}'$ satisfies $A_n \uparrow A$, then $A \in \mathscr{C}'$; if $A_1 \subset A_2$ are elements of $\mathscr{C}'$, then $A_2 \backslash A_1 \in \mathscr{C}'$. Observe also that $\mathscr{C}$ is a π-system: If $C_1, C_2 \in \mathscr{C}$, then $C_1 \cap C_2 \in \mathscr{C}$. Because $\mathscr{C} \subset \mathscr{C}'$ and the σ-algebra generated by $\mathscr{C}$ is equal to $\mathscr{B}$, it follows that $\mathscr{C}' = \mathscr{B}$ by the Dynkin $\pi\lambda$-theorem (see, for example, Billingsley [2]). Thus, (2.8) holds for all $A \in \mathscr{B}$ and $t \geq 0$, and $\zeta(\cdot)$ is a measure-valued fluid model solution.

Likewise, the previous argument shows that if $z(\cdot)$: $[0,\infty) \to \mathbb{R}_+$ is a continuous function satisfying (i) of Definition 2.1 and (2.10) for all $t \geq 0$, then $\zeta(\cdot)$ defined by (2.9) is a measure-valued fluid model solution and $z(\cdot)$ is its total mass function.

The first result establishes uniqueness of fluid model solutions under a Lipschitz assumption on the initial condition ζ_0. Note that uniqueness of fluid model solutions is equivalent to uniqueness of measure-valued fluid model solutions.

THEOREM 2.2. *Suppose that there exists a finite constant L such that*

$$\zeta_0(F\times[y,y'])\le L|y'-y| \tag{2.11}$$

for all Borel sets $F\subset\overline{\mathbb{R}}_+$ and all $y'>y\ge 0$. Then, a (measure-valued) fluid model solution for data $(\lambda,\vartheta,\zeta_0)$ is unique.

See §3 for the proof. It will be shown that the fluid model defined above describes the limit in distribution of the rescaled processes $\{\bar{\mathscr{Z}}^r(\cdot)\colon r\in\mathscr{R}\}$. The measure-valued fluid model solution $\zeta(\cdot)$ corresponds to the measure-valued state descriptor $\mathscr{Z}(\cdot)$, and the fluid model solution $z(\cdot)$ is the limit of the queue length process $Z(\cdot)$. The main result concerning the convergence of $\{\bar{\mathscr{Z}}^r(\cdot)\}$ is given below. This result also establishes the existence of (measure-valued) fluid model solutions. We first state the necessary asymptotic assumptions.

Assumptions $\mathcal{A}$. As $r\to\infty$,

$$(\bar{E}^r(t))\to(\lambda t) \tag{2.12}$$

in distribution for the topology of uniform convergence on compact sets. In particular, $\lambda^r\to\lambda$. Furthermore,

$$\breve{\vartheta}^r\xrightarrow{w}\vartheta, \tag{2.13}$$

$$\bar{\mathscr{Z}}^r(0)\xrightarrow{w}\zeta_0,\ \text{in distribution.} \tag{2.14}$$

THEOREM 2.3. *If Assumptions $\mathcal{A}$ hold, then the sequence $\{\bar{\mathscr{Z}}^r(\cdot)\colon r\in\mathscr{R}\}$ is tight and each weak limit point is almost surely a measure-valued fluid model solution $\zeta(\cdot)$ for the data $(\lambda,\vartheta,\zeta_0)$. If, in addition, Condition (2.11) holds, then $\bar{\mathscr{Z}}^r(\cdot)$ converges in distribution, as $r\to\infty$, to the unique measure-valued fluid model solution $\zeta(\cdot)$.*

The proof appears in §§5 and 6.

2.4. Properties of the fluid model. Despite the quite abstract setting of this paper (measure-valued processes), some concrete and explicit results concerning the fluid model can be obtained. Let $(\lambda,\vartheta,\zeta_0)$ satisfy the assumptions of §2.3.

The following result describes the time equilibrium of the fluid model, that is, its behavior as time tends to infinity.

THEOREM 2.4. *Suppose that $\lambda\mathbf{E}[B1_{\{D=\infty\}}]<1$ and $\mathbf{E}[\min\{B,D\}]<\infty$. Then, as $t\to\infty$, any fluid model solution $(z(t))$ converges to the unique positive solution z_∞ of the fixed point equation*

$$z_\infty=\lambda\mathbf{E}[\min\{z_\infty B,D\}].$$

Moreover, any measure-valued fluid model solution $(\zeta(t))$ converges in $\mathbf{M}$ to the unique measure ζ_∞ defined by

$$\zeta_\infty(x,y)=\lambda\int_0^\infty\mathbf{P}\left(B\ge x+\frac{t}{z_\infty};D\ge y+t\right)\mathrm{d}t=\mathbf{E}[\min\{z_\infty(B-x)^+,(D-y)^+\}] \tag{2.15}$$

for $x,y\ge 0$.

A heuristic interpretation of the important fixed point equation satisfied by z_∞ is as follows: Let Z^r denote the steady state number of jobs in the system. Furthermore, let $V^r(B)$ be the sojourn time of a job if the job never reneges. Then, the actual sojourn time is given by $\min\{V^r(B),D^r\}$, and from Little's law we get

$$\mathbf{E}[Z^r]=\lambda\mathbf{E}[\min\{V^r(B),D^r\}].$$

Divide both sides of this relation by r and let $r\to\infty$. Because we observe the system in steady state at time 0, the number of jobs hardly changes and, by a snapshot principle, we conclude that $V^r=Z^rB+\mathrm{o}(r)$. Furthermore, we have $D^r=Dr$. Noting that $Z^r/r\to z_\infty$ then gives the desired equation. In §4, this equation is analyzed to investigate the qualitative behavior of the model.

Theorems 2.2 and 2.4 are proved in §3, and Theorem 2.3 is proved in §§5 and 6. Note that ζ_∞, which describes the asymptotic behavior of the measure-valued fluid model, has a simple expression in terms of the solution z_∞ of the fixed point equation.

3. Properties of the fluid model. In this section, some basic properties of fluid model solutions are derived. In what follows, let $z(\cdot)$ be an arbitrary fluid model solution for data $(\lambda, \vartheta, \zeta_0)$. Note that existence of $z(\cdot)$ is guaranteed by Theorem 2.3, which is proved in §§5 and 6. Recall that $z(\cdot)$ satisfies (2.10):

$$z(t) = z_0\mathbf{P}(B^0 \geq S(0,t);\, D^0 \geq t) + \lambda \int_0^t \mathbf{P}(B \geq S(s,t);\, D \geq t-s)\,\mathrm{d}s.$$

If $z_0 > 0$, define

$$\tilde{S}(t) = \inf\{s\colon\, S(0,s) \geq t\}. \tag{3.1}$$

Because $z(t) \leq z_0 + \lambda t$, $S(0,t) \to \infty$ as $t \to \infty$, implying that $\tilde{S}(t)$ is well-defined for all t. In addition, $z_0 > 0$ implies that $\tilde{S}(t) < \infty$ for all t by property (i) of Definition 2.1 and continuity of $z(\cdot)$.

Define $\tilde{z}(t) = z(\tilde{S}(t))$. By a change of variables, $\tilde{S}(t) = \int_0^t \tilde{z}(u)\,\mathrm{d}u$ and $\tilde{z}(\cdot)$ satisfies the equation

$$\tilde{z}(t) = z_0\mathbf{P}(B^0 \geq t;\, D^0 \geq \tilde{S}(t)) + \lambda \int_0^t \tilde{z}(u)\mathbf{P}(B \geq t-u;\, D \geq \tilde{S}(t) - \tilde{S}(u))\,\mathrm{d}u. \tag{3.2}$$

3.1. A maximal solution. An important monotonicity property of fluid model solutions is proved in this section.

Proposition 3.1. *If* $\lambda\mathbf{E}[B1_{\{D=\infty\}}] < 1$ *and* $\mathbf{E}[\min\{B, D\}] < \infty$, *then any fluid model solution is bounded.*

Proof. Note, first, that because $\mathbf{E}[\min\{B,D\}] < \infty$, also $\mathbf{E}[\min\{aB, D\}] < \infty$ for every $a \in [0,\infty)$. Define $\|z\|_t = \sup_{0\leq u\leq t} z(u)$. Note that $\|z\|_t \leq z_0 + \lambda t < \infty$. Fix t and let $u \in [0,t]$. Because $S(s,u) \geq (u-s)/\|z\|_t$,

$$z(u) \leq z_0 + \lambda \int_0^u \mathbf{P}(\|z\|_t B \geq u-s;\, D \geq u-s)\,\mathrm{d}s \leq z_0 + \lambda\mathbf{E}[\min\{\|z\|_t B, D\}],$$

which is finite because $\mathbf{E}[\min\{B,D\}] < \infty$. By taking the supremum over $u \in [0,t]$ and by dividing both sides by $\|z\|_t$ one obtains the relation

$$1 \leq z_0/\|z\|_t + \lambda\mathbf{E}[\min\{B, D/\|z\|_t\}].$$

If $\|z\|_t \to \infty$, then by monotone convergence one gets the inequality

$$1 \leq \lambda\mathbf{E}[B1_{\{D=\infty\}}],$$

which contradicts the assumption $\lambda\mathbf{E}[B1_{\{D=\infty\}}] < 1$. We conclude that $\|z\|_t$ is bounded. □

Recall from the definition of fluid model solutions that the distributions of B^0 and D^0 are assumed to be free of atoms.

Proposition 3.2 (Maximal Solution). *There exists a fluid model solution* $z^*(\cdot)$ *for* $(\lambda, \vartheta, \zeta_0)$ *that is maximal*: *for any fluid model solution* $z(\cdot)$ *for* $(\lambda, \vartheta, \zeta_0)$, *the relation* $z(t) \leq z^*(t)$ *holds for all* $t \geq 0$.

Proof. To define $z^*(\cdot)$, we first define a sequence of functions $z^n(\cdot)$ inductively as follows. Let $z^0(t) = z_0 + \lambda t$ and, for $n \geq 0$, define $S^n(u,v) = \int_u^v (1/z^n(r))\,\mathrm{d}r$ for $v \geq u \geq 0$ and

$$z^{n+1}(t) = z_0\mathbf{P}(B^0 \geq S^n(0,t);\, D^0 \geq t) + \lambda \int_0^t \mathbf{P}(B \geq S^n(s,t);\, D \geq t-s)\,\mathrm{d}s. \tag{3.3}$$

We show that $z^{n+1}(t) \leq z^n(t)$ by induction. The inequality $z^1(t) \leq z^0(t)$ is trivial. Suppose that $z^n(t) \leq z^{n-1}(t)$. Then, $S^n(u,v) \geq S^{n-1}(u,v)$ for all $v \geq u \geq 0$ and, using the fact that tail probabilities are nonincreasing,

$$\begin{aligned} z^{n+1}(t) &= z_0\mathbf{P}(B^0 \geq S^n(0,t);\, D^0 \geq t) + \lambda \int_0^t \mathbf{P}(B \geq S^n(s,t);\, D \geq t-s)\,\mathrm{d}s \\ &\leq z_0\mathbf{P}(B^0 \geq S^{n-1}(0,t);\, D^0 \geq t) + \lambda \int_0^t \mathbf{P}(B \geq S^{n-1}(s,t);\, D \geq t-s)\,\mathrm{d}s, \end{aligned}$$

which equals $z^n(t)$.

Because $z^n(t)$ is nonincreasing in n and nonnegative for all n, there exists a function $z^*(t)$ such that $z^*(t) = \lim_{n\to\infty} z^n(t)$. If $z(\cdot)$ is any fluid model solution for $(\lambda, \vartheta, \zeta_0)$, then $z(t) \leq z^*(t)$ for all $t \geq 0$. This is true because $z(t) \leq z^0(t)$ and, using an inductive argument as above, $z(t) \leq z^n(t)$ for all n. Because we know that at least one fluid model solution exists (Theorem 2.3), it follows that $\inf_{t>a} z^*(t) > 0$ for all $a > 0$. To show that $z^*(\cdot)$ is continuous, let $t, h \geq 0$. Then,

$$|z^*(t+h) - z^*(t)| = \lim_{n\to\infty} |z^n(t+h) - z^n(t)|.$$

By definition of $z^n(\cdot)$,

$$|z^n(t+h) - z^n(t)| \leq z_0 \mathbf{P}(t \leq D^0 < t+h) + \lambda \int_0^t \mathbf{P}(t-s \leq D < t+h-s) + \lambda h.$$

The right-side tends to zero as $h \to 0$ because D^0 has no atoms and because $\int_0^\infty \mathbf{P}(w \leq D < w+h)\, dw \leq h$.

To show that $z^*(\cdot)$ satisfies (2.10), let $S^*(u, v) = \int_u^v (1/z^*(r))\, \mathrm{d}r$ for all $v \geq u \geq 0$. Then, $S^n(s,t) \to S^*(s,t)$ for all $t \geq s \geq 0$ by monotone convergence. Because $z^*(\cdot)$ is bounded away from zero on $(0, \infty)$, $S^*(s,t)$ is strictly decreasing in s. Thus, because B and D have at most countably many atoms, there are at most countably many s such that $S^*(s,t)$ is an atom of B or $t-s$ is an atom of D. This implies that $\mathbf{P}(B \geq S^n(s,t); D \geq t-s) \to \mathbf{P}(B \geq S^*(s,t); D \geq t-s)$ for almost every $s \in [0,t]$. Take the limit as $n \to \infty$ in (3.3). By the previous discussion, the integral term converges to $\lambda \int_0^t \mathbf{P}(B \geq S^*(s,t); D \geq t-s)\, \mathrm{d}s$ by bounded convergence. Because B^0 and D^0 have no atoms, the first right-hand term converges to $z_0 \mathbf{P}(B^0 \geq S^*(0,t); D^0 \geq t)$.

We conclude that $z^*(t)$ is a maximal fluid model solution because it is continuous and satisfies (2.10) and (i) of Definition 2.1. □

3.2. Convergence of fluid model solutions. In this subsection, we show the convergence of fluid model solutions to a nontrivial constant z_∞ as $t \to \infty$.

Proposition 3.3. *If $\lambda \mathbf{E}[B 1_{\{D=\infty\}}] < 1$, $\mathbf{E}[\min\{B, D\}] < \infty$, and $\rho > 1$, then the equation*

$$z_\infty = \lambda \mathbf{E}[\min\{z_\infty B, D\}] \tag{3.4}$$

has a unique solution in $(0, \infty)$.

Proof. The function $f\colon a \mapsto \lambda \mathbf{E}[\min\{B, aD\}]$ is nondecreasing and concave on $[0, \infty)$. Note that $f(a) = \lambda \mathbf{E}[\min\{B, aD\} 1_{\{D<\infty\}}] + \lambda \mathbf{E}[B 1_{\{D=\infty\}}]$ for $a > 0$. Therefore, f is continuous on $(0, \infty)$, $\lim_{a\to 0} f(a) = \lambda \mathbf{E}[B 1_{\{D=\infty\}}] < 1$, and $\lim_{a\to\infty} f(a) = \lambda \mathbf{E}[B] > 1$. Thus, there exists $a_0 \in (0,\infty)$ such that $f(a_0) = 1$. Concavity and monotonicity imply that a_0 is unique. Otherwise, f would be constant and equal to 1 after a_0, which contradicts $\lim_{a\to\infty} f(a) = \lambda \mathbf{E}[B] > 1$. We conclude that $1/a_0$ is the unique solution of (3.4). □

We are now ready to present the main result of this subsection, concerning the asymptotic behavior of any fluid model solution $(z(t))$ as t goes to infinity.

Theorem 3.4. *Let $z(\cdot)$ be a fluid model solution for $(\lambda, \vartheta, \zeta_0)$ and assume that*

$$\lambda \mathbf{E}[B 1_{\{D=\infty\}}] < 1, \qquad \mathbf{E}[\min\{B, D\}] < \infty, \qquad \textit{and} \qquad \rho > 1.$$

Then, as $t \to \infty$, $z(t)$ converges to z_∞, the unique positive solution of the fixed point (3.4).

Proof. It suffices to show that $\bar{z} = \limsup_{t\to\infty} z(t) \leq z_\infty$ and $\underline{z} = \liminf_{t\to\infty} z(t) \geq z_\infty$. We start with the former. Proposition 3.1 implies that $\bar{z} < \infty$. For any $\varepsilon > 0$, there exists a t_ε such that $t > t_\varepsilon$ implies $z(t) \leq \bar{z} + \varepsilon$. So, for $t > t_\varepsilon$, (2.10) yields

$$z(t) \leq z_0 \mathbf{P}(B^0 \geq S(0,t); D^0 \geq t) + \lambda \int_0^{t_\varepsilon} \mathbf{P}(B \geq S(s,t); D \geq t-s)\, \mathrm{d}s$$
$$+ \lambda \int_0^{t-t_\varepsilon} \mathbf{P}(\min\{(\bar{z}+\varepsilon)B, D\} \geq s)\, \mathrm{d}s.$$

Hence, taking the lim sup on both sides and noting that $\lim_{t\to\infty} S(s,t) = \infty$ for each fixed $s \geq 0$ yields

$$\bar{z} \leq \lambda \mathbf{E}[\min\{(\bar{z}+\varepsilon)B, D\}].$$

Letting $\varepsilon \downarrow 0$, we obtain $\bar{z} \leq z_\infty$ by the dominated convergence theorem. The lower bound follows by an analogous argument, after first noting that $\underline{z} > 0$ because $z(\cdot)$ is a fluid model solution. □

3.3. Uniqueness of fluid model solutions under nonzero initial conditions. The uniqueness of fluid model solutions is difficult to determine in general. If one looks at the time-changed Equation (3.2) and takes $\lambda = 0$, one gets an ordinary differential equation (ODE). Uniqueness of solutions to such an ODE can usually only be established by reducing it to some special case or by imposing a Lipschitz condition. If $D \equiv \infty$, then (3.2) reduces to a renewal equation for which uniqueness is known to hold.

Unfortunately, this reduction is not possible for more general D. For technical reasons, one has to use a Lipschitz condition on the distribution function of (B^0, D^0). This condition is probably stronger than necessary, so that the question of uniqueness in general is still open. For related results in the functional analysis literature, we refer to Chapter 2 of Hale and Verduyn Lunel [15].

THEOREM 3.5. *Let $(\lambda, \vartheta, \zeta_0)$ be such that $z_0 > 0$ and $F_0(x, y) = \mathbf{P}(B^0 \geq x; D^0 \geq y)$ is Lipschitz continuous in y, that is, there is a finite constant L such that for all $x, y, y' \in \overline{\mathbb{R}}_+$,*

$$|F_0(x, y) - F_0(x, y')| \leq L|y - y'|.$$

Then, a fluid model solution for $(\lambda, \vartheta, \zeta_0)$ is unique.

Let $z(\cdot)$ be a fluid model solution for $(\lambda, \vartheta, \zeta_0)$ and let $\zeta(\cdot)$ be the corresponding measure-valued fluid model solution. Let $\tilde{S}(\cdot)$ be defined as in (3.1) and define $\tilde{z}(t) = z(\tilde{S}(t))$. Recall that $\tilde{z}(\cdot)$ satisfies (3.2). Let $\tilde{\zeta}(t) = \zeta(\tilde{S}(t))$. It can easily be shown that for all $u, v \in \overline{\mathbb{R}}_+$,

$$\tilde{\zeta}(t)(u, v) = z_0 \mathbf{P}(B^0 \geq u + t; D^0 \geq v + \tilde{S}(t)) + \lambda \int_0^t \tilde{z}(s) \mathbf{P}(B \geq u + (t - s); D \geq v + \tilde{S}(t) - \tilde{S}(s))\, \mathrm{d}s. \tag{3.5}$$

Clearly, for any $u, v \in \overline{\mathbb{R}}_+$, and $t \geq 0$, $\tilde{\zeta}(t)(u, v)$ is completely determined by $\tilde{z}(t)$ and $(\lambda, \vartheta, \zeta_0)$. Thus, uniqueness of $\tilde{z}(t)$ on a time interval A implies uniqueness of $\tilde{\zeta}(t)$ on A.

We next introduce *shifted* time-changed fluid model solutions. For $t_0, t \geq 0$, define $\tilde{z}(t_0, t) = \tilde{z}(t_0 + t)$ and

$$\tilde{S}(t_0, u, v) = \int_u^v \tilde{z}(t_0, s)\, \mathrm{d}s = \tilde{S}(t_0 + v) - \tilde{S}(t_0 + u).$$

It is easily checked that for all $t_0, t \geq 0$,

$$\tilde{z}(t_0, t) = \tilde{\zeta}(t_0)(t, \tilde{S}(t_0, 0, t)) + \lambda \int_0^t \tilde{z}(t_0, s) \mathbf{P}(B \geq t - s; D \geq \tilde{S}(t_0, s, t))\, \mathrm{d}s. \tag{3.6}$$

The idea of the proof is as follows: We take a suitable constant $a > 0$ and prove first that $\tilde{z}(t)$ is unique on $[0, a]$. As indicated above, uniqueness carries over to $\tilde{\zeta}(t)$ for $t \in [0, a]$. Using this and the shifted Equation (3.6), we then prove uniqueness for $\tilde{z}(t)$ on the interval $[a, 2a]$ and so forth. This iterative procedure works if $\tilde{\zeta}(t)(u, v)$ is Lipschitz in v for all $t \geq 0$. This is the content of the following lemma.

LEMMA 3.6. *Under the assumptions of Theorem 3.5, we have for all $x, y, y' \in \overline{\mathbb{R}}_+$, and $t \geq 0$,*

$$|\tilde{\zeta}(t)(x, y) - \tilde{\zeta}(t)(x, y')| \leq (z_0 L + \lambda)|y - y'|.$$

PROOF. We may assume that $y \leq y'$. From (3.5) we obtain

$$|\tilde{\zeta}(t)(x, y) - \tilde{\zeta}(t)(x, y')| \leq z_0 L|y' - y| + \lambda \int_0^t \tilde{z}(s) \mathbf{P}(\tilde{S}(t) - \tilde{S}(s) + y \leq D < \tilde{S}(t) - \tilde{S}(s) + y')\, \mathrm{d}s.$$

Noting that $\tilde{z}(s)\mathrm{d}s = \mathrm{d}\tilde{S}(s)$, we can rewrite the right side as

$$z_0 L|y' - y| + \lambda \int_0^{\tilde{S}(t)} \mathbf{P}(r + y \leq D < r + y')\, \mathrm{d}r. \tag{3.7}$$

Because for any $\delta > 0$,

$$\int_0^\infty \mathbf{P}(w \leq D < w + \delta)\, \mathrm{d}w \leq \delta,$$

we see that (3.7) can be bounded above by $(z_0 L + \lambda)|y - y'|$. $\square$

Proof of Theorem 3.5. Recall that we have fixed a fluid model solution $z(\cdot)$. By the one-to-one correspondence between solutions of (2.10) and (3.2), it suffices to show that $\tilde{z}(\cdot)$ is unique. Let $a = 1/(2(z_0 L + 4\lambda))$. We first show that $\tilde{z}(\cdot)$ is unique on $[0, a]$. For this, suppose that there exists a continuous function $h(\cdot)$ that satisfies (3.2) for all $t \in [0, a]$, with $H(t) = \int_0^t h(s)\,\mathrm{d}s$ in place of $\tilde{S}(\cdot)$. Set $\varepsilon = \sup_{0\le t\le a} |\tilde{z}(t) - h(t)|$. Then, for all $0 \le s < t \le a$,

$$|H(t) - H(s) - (\tilde{S}(t) - \tilde{S}(s))| \le \varepsilon a.$$

Recall that $\tilde{S}(t) = \int_0^t \tilde{z}(s)\,\mathrm{d}s$. Using (3.2) for both $z(\cdot)$ and $h(\cdot)$ together with the Lipschitz assumption, we obtain for $t \in [0, a]$,

$$\begin{aligned}|\tilde{z}(t) - h(t)| \le{} & z_0 |\mathbf{P}(B^0 \ge t;\, D^0 \ge \tilde{S}(t)) - \mathbf{P}(B^0 \ge t;\, D^0 \ge H(t))| \\ & + \lambda \int_0^t |\tilde{z}(s)\mathbf{P}(B \ge t-s;\, D \ge \tilde{S}(t) - \tilde{S}(s)) - h(s)\mathbf{P}(B \ge t-s;\, D \ge H(t) - H(s))|\,\mathrm{d}s.\end{aligned}$$

The first right-hand term is bounded by $z_0 L\varepsilon a$. The second right-hand term is bounded by

$$\lambda \int_0^t |\tilde{z}(s) - h(s)|\,\mathrm{d}s + \lambda \int_0^t \tilde{z}(s)|\mathbf{P}(B \ge t-s;\, D \ge \tilde{S}(t) - \tilde{S}(s)) - \mathbf{P}(B \ge t-s;\, D \ge H(t) - H(s))|\,\mathrm{d}s.$$

Call the previous two terms IIa and IIb. We have $IIa \le \lambda\varepsilon a$. To bound IIb, we use the inequality

$$|\mathbf{P}(B \ge t-s; D \ge \tilde{S}(t) - \tilde{S}(s)) - \mathbf{P}(B \ge t-s; D \ge H(t) - H(s))| \le \mathbf{P}(\tilde{S}(t) - \tilde{S}(s) - \varepsilon a \le D < \tilde{S}(t) - \tilde{S}(s) + \varepsilon a)$$

to obtain (after a change of variables $r = \tilde{S}(t) - \tilde{S}(s)$)

$$IIb \le \lambda \int_0^{\tilde{S}(t)} \mathbf{P}(r - \varepsilon a \le D < r + \varepsilon a)\,\mathrm{d}r \le \lambda 2\varepsilon a.$$

Putting everything together, we see that for $t \in [0, a]$,

$$|\tilde{z}(t) - h(t)| \le z_0 L\varepsilon a + 3\lambda\varepsilon a \le \varepsilon/2,$$

which implies that $\varepsilon = 0$. Hence, $\tilde{z}(\cdot)$ and $h(\cdot)$ coincide on $[0, a]$, and so $\tilde{z}(\cdot)$ is unique on $[0, a]$.

Suppose now that $\tilde{z}(\cdot)$ is unique on $[0, ka]$ for some $k \ge 1$. This uniquely determines $\tilde{\zeta}(ka)$. By (3.6), $\tilde{z}(\cdot)$ satisfies

$$\tilde{z}(ka, t) = \tilde{\zeta}(ka)(t, \tilde{S}(ka, 0, t)) + \lambda \int_0^t \tilde{z}(ka, s)\mathbf{P}(B \ge t-s;\, D \ge \tilde{S}(ka+t) - \tilde{S}(ka+s))\,\mathrm{d}s, \quad t \ge 0. \tag{3.8}$$

We now show that this shifted equation has a unique solution on $[0, a]$, implying that $\tilde{z}(\cdot)$ is unique on $[0, (k+1)a]$. Suppose $h(\cdot)$ also satisfies (3.8), with $H(u, v) = \int_u^v h(s)\,\mathrm{d}s$ in place of $\tilde{S}(ka, u, v)$ for all $v \ge u \ge 0$. Set $\varepsilon = \sup_{t\in[0,a]} |\tilde{z}(ka, t) - h(t)|$. As before, we have

$$|(\tilde{S}(ka, t) - \tilde{S}(ka, s)) - (H(t) - H(s))| \le (t-s)\varepsilon \le a\varepsilon.$$

Using this, we get as before (this time using Lemma 3.6 for the first term on the right)

$$|\tilde{z}(ka, t) - h(t)| \le (z_0 L + \lambda)a\varepsilon + 3\lambda a\varepsilon \le \varepsilon/2.$$

Thus, $\varepsilon = 0$ and $\tilde{z}(\cdot)$ is unique on $[0, (k+1)a]$. Iterating this argument completes the proof for all $t \ge 0$. □

3.4. Uniqueness starting from zero. The result in this subsection can be seen as an extension of a result in Puha et al. [24], where the case of a PS queue without impatience ($D \equiv \infty$) was considered.

Theorem 3.7. *Suppose that for each $\varepsilon > 0$, there is a nonincreasing (in each coordinate) function $F_\varepsilon\colon \overline{\mathbb{R}}_+^2 \to \mathbb{R}$, with $F_\varepsilon(0,0) > 0$ and $0 \le F_\varepsilon(x, y) \le \lambda\varepsilon$, such that the equation*

$$z_\varepsilon(t) = F_\varepsilon(S_\varepsilon(0, t), t) + \lambda \int_0^t \mathbf{P}(B \ge S_\varepsilon(t-s, t);\, D \ge s)\,\mathrm{d}s,$$

where $S_\varepsilon(s, t) = \int_s^t 1/z_\varepsilon(u)\,\mathrm{d}u$ has a unique solution $z_\varepsilon(\cdot)$ satisfying $\inf_{t>a} z_\varepsilon(t) > 0$ for all $a > 0$. Then, for each $t \ge 0$, $z_\varepsilon(t) \to z_0^*(t)$ as $\varepsilon \downarrow 0$, where $z_0^*(\cdot)$ is the maximal solution starting from zero.

Proof. As in the proof of Proposition 3.2, $z_\varepsilon(\cdot)$ can be written as the pointwise limit $\lim_{n\to\infty} z_\varepsilon^n(\cdot)$, with $z_\varepsilon^n(\cdot)$ recursively defined by $z_\varepsilon^0(t) = F_\varepsilon(0,0) + \lambda t$ and

$$z_\varepsilon^{n+1}(t) = F_\varepsilon(S_\varepsilon^n(0,t),t) + \lambda \int_0^t \mathbf{P}(B \ge S_\varepsilon^n(t-s,t); D \ge s)\,\mathrm{d}s.$$

From this construction, it is easily shown by induction on n that $z_\varepsilon^n(t)$ is nonincreasing in n and that $z_\varepsilon^n(t) \ge z_0^n(t)$. Induction on n also yields

$$\limsup_{\varepsilon\downarrow 0} z_\varepsilon(t) \le \limsup_{\varepsilon\downarrow 0} z_\varepsilon^n(t) = z_0^n(t).$$

Because this holds for any n, and $z_0^n(t) \to z_0^*(t)$, we can let $n\to\infty$ to obtain

$$\limsup_{\varepsilon\downarrow 0} z_\varepsilon(t) \le z_0^*(t).$$

Similarly, for every $n \ge 0$,

$$z_0^*(t) = \lim_{n\to\infty} z_0^n(t) \le \limsup_{n\to\infty} z_\varepsilon^n(t) = z_\varepsilon(t).$$

We conclude that $z_\varepsilon(t) \ge z_0^*(t)$ for every $\varepsilon > 0$, which implies the lower limit and the convergence $z_\varepsilon(t) \to z_0^*(t)$. □

Uniqueness of fluid model solutions starting from zero is now a simple corollary.

Corollary 3.8. *A fluid model solution starting from zero is unique.*

Proof. Let $z(\cdot)$ be a fluid model solution starting from zero. Define $z_\varepsilon(t) = z(t+\varepsilon)$. Then, $z_\varepsilon(\cdot)$ satisfies the equation

$$z_\varepsilon(t) = F_\varepsilon(S_\varepsilon(0,t),t) + \lambda \int_0^t \mathbf{P}(B \ge S_\varepsilon(t-s,t); D \ge s)\,\mathrm{d}s, \tag{3.9}$$

where

$$F_\varepsilon(x,y) = \lambda \int_0^\varepsilon \mathbf{P}\left(B \ge x + \int_{\varepsilon-s}^{\varepsilon} \frac{1}{z(u)}\,\mathrm{d}u; D \ge s+y\right)\mathrm{d}s.$$

Observe that for all $\varepsilon > 0$, F_ε satisfies the assumptions of Theorem 3.7. Moreover, F_ε is globally Lipschitz in the second coordinate (with Lipschitz constant 1). Let $(B_\varepsilon^0, D_\varepsilon^0)$ be distributed as $F_\varepsilon(\cdot,\cdot)/F_\varepsilon(0,0)$. Then, by Theorem 3.5, with $(B_\varepsilon^0, D_\varepsilon^0)$ in place of (B^0, D^0), (3.9) has a unique solution so that $z_\varepsilon(\cdot)$ is uniquely determined by $z(t), 0 \le t \le \varepsilon$. Because $F_\varepsilon(x,y) \le \lambda\varepsilon$, we see from the previous theorem that $z_\varepsilon(t) \to z_0^*(t)$ as $\varepsilon \to 0$. Also, $z_\varepsilon(t) = z(t+\varepsilon) \to z(t)$, because $z(t)$ is continuous. We conclude that $z(t) = z_0^*(t)$, which implies uniqueness. □

3.5. Analysis of the measure-valued fluid model. Some properties of measure-valued fluid model solutions, which are analogues of properties of fluid model solutions, are gathered in the next theorem.

Theorem 3.9. *Let $\zeta(\cdot)$ be a measure-valued fluid model solution for $(\lambda, \vartheta, \zeta_0)$ with total mass function $z(\cdot)$.*

(i) *Suppose that $\rho > 1$, $\mathbf{E}[\min\{B,D\}] < \infty$, and $\lambda\mathbf{E}[B1_{\{D=\infty\}}] < 1$. Then, $\zeta(t) \xrightarrow{\mathrm{w}} \zeta_{z_\infty}$ as $t\to\infty$, where for each $c > 0$, the measure ζ_c is defined by*

$$\zeta_c([x,\infty]\times[y,\infty]) = \lambda \int_0^\infty \mathbf{P}(B \ge x + sc^{-1}, D \ge y+s)\,\mathrm{d}s, \quad x,y \in \overline{\mathbb{R}}_+,$$

and z_∞ is the unique positive solution of the fixed point Equation (3.4).

(ii) *If Condition (2.11) of Theorem 2.2 holds, then $\zeta(\cdot)$ is unique.*

Proof. By Theorem 3.4, $z(t) \to z_\infty$ as $t\to\infty$. Continuity of $z(\cdot)$ and finiteness of z_∞ imply that there exists an $M < \infty$ such that $z(t) \le M$ for all $t \ge 0$. This implies that $S(s,t) \ge (t-s)M^{-1}$ for all $s,t \ge 0$. We first show that $\{\zeta(t)\colon t \ge 0\}$ is relatively compact in $\mathbf{M}$. For $n \in \mathbb{N}$, let

$$A_n = \{[0,n]\times[0,n]\} \cup \{\infty\times[0,n]\} \cup \{[0,n]\times\infty\} \cup \{(\infty,\infty)\}.$$

Clearly, $A_n \uparrow \overline{\mathbb{R}}_+^2$ and $A_n^c + (x,y) \subset A_n^c + (x',y')$ for all $x \ge x'$ and $y \ge y'$, where A_n^c denotes the complement. By (2.8), for all $t \ge 0$ and $n \in \mathbb{N}$,

$$\begin{aligned}\zeta(t)(A_n^c) &= \zeta_0(A_n^c + (S(0,t),t)) + \lambda \int_0^t \vartheta(A_n^c + (S(s,t),t-s))\,\mathrm{d}s \\ &\le \zeta_0(A_n^c) + \lambda \int_0^\infty \vartheta(A_n^c + (sM^{-1},s))\,\mathrm{d}s \\ &= \zeta_0(A_n^c) + \zeta_M(A_n^c).\end{aligned}$$

Thus, $\lim_{n\to\infty}\sup_{t\ge 0}\zeta(t)(A_n^c)=0$. Because $\sup_{t\ge 0}\zeta(t)(\overline{\mathbb{R}}_+^2)=\sup_{t\ge 0} z(t)\le M$, the set $\{\zeta(t)\colon t\ge 0\}$ is relatively compact in $\mathbf{M}$ (see Kallenberg [17, Theorem A 7.5]). Let $\zeta^*\in\mathbf{M}$ be a weak limit point (along a subsequence) of $\zeta(t)$ as $t\to\infty$, and let $C\in\mathscr{C}$ be a continuity set of ζ^*. Then, $\zeta(t)(C)\to\zeta^*(C)$. Because $z(t)\to z_\infty$, $S(t-t_0,t)\to t_0/z_\infty$ for each $t_0>0$. For all $0\le t_0<t$,

$$\zeta(t)(C)=\zeta_0(C+(S(0,t),t))+\lambda\int_0^{t_0}\vartheta(C+(S(t-s,t),s))\,\mathrm{d}s$$
$$+\lambda\int_{t_0}^{t}\vartheta(C+(S(t-s,t),s))\,\mathrm{d}s.$$

Denote the three right-hand terms by I, II, III. Because $z(t)\to z_\infty>0$, $(S(0,t),t)\to(\infty,\infty)$, and so the first term converges to zero. Because $C\subset\overline{\mathbb{R}}_2^+$,

$$III\le\lambda\int_{t_0}^{t}\mathbf{P}(B\ge sM^{-1};\,D\ge s)\,\mathrm{d}s.$$

From this bound, it follows that $III\to 0$ as $t_0\to\infty$. Because the marginals of ϑ have at most countably many atoms, $C+(s/z_\infty,s)$ is a continuity set for ϑ for almost every s. Because $S(t-s,t)\to s/z_\infty$ on $[0,t_0]$ as $t\to\infty$, the bounded convergence theorem implies that

$$II\to\lambda\int_0^{t_0}\vartheta(C+(s/z_\infty,s))\,\mathrm{d}s,\quad\text{as } t\to\infty.$$

So we can take $t\to\infty$ and then $t_0\to\infty$ to conclude that $\zeta(t)(C)\to\zeta_{z_\infty}(C)$ as $t\to\infty$. This implies that $\zeta^*(C)=\zeta_{z_\infty}(C)$ for all $C\in\mathscr{C}$ that are ζ^*-continuity sets, which excludes at most countably many $C\in\mathscr{C}$ (because the marginals of ζ^* have at most countably many atoms). Thus, $\zeta^*(C)=\zeta_{z_\infty}(C)$ for all $C\in\mathscr{C}$, and so $\zeta^*=\zeta_{z_\infty}$ and $\zeta(t)\xrightarrow{w}\zeta_{z_\infty}$. This proves (i).

To prove (ii), note that $z(\cdot)$ is unique by Theorem 3.5 and Corollary 3.8. As discussed in §2.3, uniqueness of $\zeta(\cdot)$ follows. □

4. Applications. In this section, we analyze a number of quantitative properties of the fluid model Equation (2.10). In particular, we investigate the fixed point equation

$$z_\infty=\lambda\mathbf{E}[\min\{z_\infty B,D\}].\tag{4.1}$$

We treat a number of examples which allow for explicit computations, and also obtain a number of stochastic ordering results. In addition, we investigate the time-dependent behavior of $z(t)$ for exponentially distributed lead times.

Apart from the mean queue length z, we are also interested in the long-term fraction of jobs that leave the system after successful service completion. Denote this fraction by P_s. It is clear that $P_s=\mathbf{P}(D>z_\infty B)$. Consider two systems indexed by 1 and 2 such that $(B_2,D_2)\equiv(B_1,aD_1)$ for some $a>0$, and such that $\lambda_1=\lambda_2$. Then, (with obvious notation) we have

$$z_{2,\infty}=az_{1,\infty},\qquad P_{s,2}=P_{s,1}.$$

Consequently, the fraction P_s is invariant under any rescaling of D.

We now proceed by analyzing a number of special cases. In §4.1, we assume a strong form of dependence. Section 4.3 assumes that B and D are independent. We give a remarkably simple expression for $z(t)$ in the case that D has an exponential distribution. Finally, §4.3 considers an example which can be used as a flow level model for the integration of elastic and streaming traffic.

4.1. Completely dependent lead times. Consider first the case $D=\Theta B$, where $\Theta>0$ is a random variable (independent of B) reflecting the average service rate expected by a job. In this case, the performance measures can be determined from the equations (recall that $\rho=\lambda\mathbf{E}[B]>1$)

$$z_\infty=\rho\mathbf{E}[\min\{\Theta,z_\infty\}],\qquad P_s=\mathbf{P}(\Theta>z_\infty).$$

Some specific examples:

—Θ *single valued.* If we assume that $\Theta=\theta$, then $z_\infty=\rho\min\{\theta,z_\infty\}$, which implies that $z_\infty=\rho\theta$ because $\rho>1$. From this, it follows that all jobs leave the system impatiently: $P_s=\mathbf{P}(\theta>\rho\theta)=0$. Observe that when a job leaves the system, a fraction $1/\rho$ of its service time has been processed.

—Θ *two valued.* From the previous example, it is clear that $P_s > 0$ only if some jobs are more patient than others. In this example, we assume that Θ equals θ_1 with probability p and θ_2 with probability $1-p$. Take $\theta_2 > \theta_1$. Equation (4.1) now simplifies to

$$z_\infty = \rho p \min\{z_\infty, \theta_1\} + \rho(1-p)\min\{z_\infty, \theta_2\}.$$

From this equation and the properties $\theta_2 > \theta_1, \rho > 1$, it follows that $z_\infty > \theta_1$. Furthermore, $z_\infty < \theta_2$ holds if and only if the equation

$$z_\infty = \rho p \theta_1 + \rho(1-p) z_\infty$$

has a nonnegative solution, which is the case if and only if $\rho(1-p) < 1$ (i.e., when the offered load of the more patient jobs does not saturate the system). In this case, we have

$$z_\infty = \frac{\rho p \theta_1}{1-\rho(1-p)} < \theta_2.$$

If the last inequality is not valid or if $\rho(1-p) \geq 1$, we must have $z_\infty \geq \theta_2$ which implies

$$z_\infty = \rho p \theta_1 + \rho(1-p)\theta_2.$$

From the above, we can conclude that $P_s = 0$ if and only if $(1-\rho(1-p))\theta_2 \leq \rho p \theta_1$. If the reverse inequality holds, then all of the more patient jobs are being served successfully, i.e., $P_s = (1-p)$.

—Θ *exponentially distributed.* Assume that the mean of Θ equals 1. In this case, z_∞ can be determined from the equation $z_\infty = \rho(1-(z_\infty+1)\mathrm{e}^{-z_\infty})$ and $P_s = \mathrm{e}^{-z_\infty}$.

Because P_s does not depend on the mean of Θ and because constant Θ yields the worst case $P_s = 0$, it seems natural to conjecture that the fraction of successful completions is positively correlated to the variability of Θ. Thus, it seems worthwhile to look for ordering relations for P_s if $\Theta_1 \overset{cvx}{\geq} \Theta_2$. If $\mathbf{E}[\Theta_1] = \mathbf{E}[\Theta_2]$, this is equivalent to $\mathbf{E}[\min\{x, \Theta_1\}] \leq \mathbf{E}[\min\{x, \Theta_2\}]$ for all $x \geq 0$. So $\Theta_1 \overset{cvx}{\geq} \Theta_2$ implies that $z_{2,\infty} \geq z_{1,\infty}$, that is, less variability in reneging behavior implies a lower service rate. To prove that $\mathbf{P}(\Theta_1 > z_{1,\infty}) \geq \mathbf{P}(\Theta_2 > z_{2,\infty})$ as well seems difficult without imposing further assumptions.

4.2. Independent lead times. In this case, we can write (4.1) as

$$\lambda \int_0^\infty \mathbf{P}(B > u)\mathbf{P}(D > z_\infty u)\,\mathrm{d}u = 1,$$

which, in case $\mathbf{E}[B] < \infty$, is equivalent to $\mathbf{P}(D > z_\infty B^*) = 1/\rho$, with B^* a random variable with density $\mathbf{P}(B > x)/\mathbf{E}[B]$ which is also independent of D.

Recall that $P_s = \mathbf{P}(D > z_\infty B)$. Consequently, if B is exponentially distributed, we have the insensitivity result (with respect to the distribution of D) $P_s = 1/\rho$. The inequality $P_s \leq 1/\rho$ holds if B^* is stochastically dominated by B, and $P_s \geq 1/\rho$ if the opposite is true. Because B^* being stochastically dominated by B is related to low variability of B, we see again that more variability (this time in the service times) leads to a better system performance (i.e., higher P_s).

Exponential reneging. If we assume that D has an exponential distribution with parameter ν (and B a general distribution), we see that z_∞ is the solution of

$$\rho\beta^*(z_\infty \nu) = 1, \tag{4.2}$$

with $\beta^*(s) = \mathbf{E}[\mathrm{e}^{-sB^*}]$. In addition, we have the following remarkable expression for the complete fluid limit $z(\cdot)$, if $z_0 = 0$:

Proposition 4.1. *Suppose $\mathbf{P}(D > t) = \mathrm{e}^{-\nu t}$, that B is independent of D, and that $z_0 = 0$. Then, the unique fluid model solution is given by*

$$z(t) = z_\infty(1-\mathrm{e}^{-\nu t}), \tag{4.3}$$

with z_∞ the solution of (4.2).

Proof. Recall that (2.10) has a unique solution. We show that (4.3) is indeed the solution of (2.10) by verification. We thus compute the right-hand side of (2.10), writing $z(u) = z_\infty(1-\mathrm{e}^{-\nu u})$.

Observe that

$$z_\infty \int_s^t \frac{1}{z(u)}\,\mathrm{d}u = \frac{1}{\nu}(\log(\mathrm{e}^{\nu t}-1) - \log(\mathrm{e}^{\nu s}-1)).$$

Consequently,

$$\begin{aligned}\lambda\int_0^t \mathbf{P}(D\ge t-s)\mathbf{P}\left(B\ge \int_s^t (1/z(u))\,\mathrm{d}u\right)\mathrm{d}s &= \frac{\lambda}{\nu}\mathrm{e}^{-\nu t}\int_0^t \mathbf{P}(z_\infty \nu B\ge \log(\mathrm{e}^{\nu t}-1)-\log(\mathrm{e}^{\nu s}-1))\,d\mathrm{e}^{\nu s}\\ &= \frac{\lambda}{\nu}\mathrm{e}^{-\nu t}\int_{-\log(\mathrm{e}^{\nu t}-1)}^\infty \mathrm{e}^{-v}\mathbf{P}(z_\infty\nu B\ge \log(\mathrm{e}^{\nu t}-1)+v)\,\mathrm{d}v\\ &= \frac{\lambda}{\nu}\mathrm{e}^{-\nu t}(\mathrm{e}^{\nu t}-1)\int_0^\infty \mathbf{P}(z_\infty\nu B\ge v)\mathrm{e}^{-v}\,\mathrm{d}v\\ &= z_\infty(1-\mathrm{e}^{-\nu t})\rho\beta^*(z_\infty\nu) = z_\infty(1-\mathrm{e}^{-\nu t}),\end{aligned}$$

which shows that $z_\infty(1-\mathrm{e}^{-\nu t})$ satisfies (2.10). ☐

4.3. TCP-friendly traffic. Assume that there exist independent random variables B_1 and D_1, with finite means, such that

$$(B,D) = \begin{cases}(B_1,\infty), & \text{with probability } p,\\ (\infty, D_1), & \text{with probability } 1-p.\end{cases}$$

When we view PS as a way of modeling TCP, this example models the integration of elastic (TCP) traffic and TCP-friendly user datagram protocol (UDP) traffic; see Key et al. [19] for a related model. The latter type of traffic is using the system for a certain amount of time, regardless of the level of congestion.

The fixed point Equation (4.1) specializes to

$$z_\infty = \lambda p\mathbf{E}[z_\infty B_1] + \lambda(1-p)\mathbf{E}[D_1].$$

Consequently, if the stability condition $\lambda p\mathbf{E}[B_1] < 1$ is satisfied, we see that

$$z_\infty = \frac{\lambda(1-p)\mathbf{E}[D_1]}{1-\lambda p\mathbf{E}[B_1]}.$$

5. Tightness. In this section, we prove the first part of Theorem 2.3. That is, we show that the sequence of processes $\{\bar{\mathcal{Z}}^r(\cdot), r\in\mathcal{R}\}$ is tight in $\mathbf{D}([0,\infty),\mathbf{M})$. The main results in this section implying this property are the compact containment result in Lemma 5.2 and an oscillation inequality in Lemma 5.6. To prove these results, a number of further lemmas are developed. Section 5.1 derives a Glivenko-Cantelli theorem for the stochastic primitives. Section 5.2 introduces a fluid scaled version of the dynamic equation for $\bar{\mathcal{Z}}^r(\cdot)$. The compact containment property is derived in §5.3. Section 5.4 serves as a preparation for the oscillation bound. In particular, it is shown that $\bar{\mathcal{Z}}^r(t)$ charges arbitrarily small mass to thin L-shaped sets. The oscillation bound is then shown in §5.5.

Throughout this section, it is assumed that Assumptions $\mathcal{A}$ hold.

5.1. A Glivenko-Cantelli theorem. An important preliminary result is the following functional Glivenko-Cantelli theorem for the stochastic primitives. It will be used in §§5.3–5.5. It is convenient to consider the primitives together as a single, measure-valued arrival process. For $r\in\mathcal{R}$ and $t\ge s\ge 0$, define the fluid scaled measure-valued arrival process by

$$\bar{\mathcal{L}}^r(t) = \frac{1}{r}\sum_{i=1}^{r\bar{E}^r(t)} \delta_{(B_i^r, D_i^r r^{-1})},$$

and define the fluid scaled increment

$$\bar{\mathcal{L}}^r(s,t) = \bar{\mathcal{L}}^r(t) - \bar{\mathcal{L}}^r(s). \tag{5.1}$$

Note that $\bar{\mathcal{L}}^r(\cdot)$ is a random element of $\mathbf{D}([0,\infty),\mathbf{M})$ and for each $t\ge s\ge 0$, $\bar{\mathcal{L}}^r(s,t)$ is a random element of $\mathbf{M}$.

To state and prove the result, we first introduce some notions from empirical process theory. Our primary reference is van der Vaart and Wellner [28]. A collection $\mathscr{C}$ of subsets of $\overline{\mathbb{R}}_+^2$ *shatters* an n-point subset $\{x_1, \ldots, x_n\} \subset \overline{\mathbb{R}}_+^2$ if the collection $\{C \cap \{x_1, \ldots, x_n\}: C \in \mathscr{C}\}$ has cardinality 2^n. In this case, say that $\mathscr{C}$ *picks out* all subsets of $\{x_1, \ldots, x_n\}$. The *Vapnik-Červonenkis index* (*VC-index*) of $\mathscr{C}$ is

$$V_{\mathscr{C}} = \min\{n: \mathscr{C} \text{ shatters no } n\text{-point subset}\},$$

where the minimum of the empty set equals infinity. The collection $\mathscr{C}$ is a *Vapnik-Červonenkis class* (*VC-class*) if it has finite VC-index.

VC-classes satisfy a useful entropy bound. Let $\mathscr{Q}$ denote the set of Borel probability measures on $\overline{\mathbb{R}}_+^2$ and, for $Q \in \mathscr{Q}$, let $\|f\|_Q = \langle |f|, Q\rangle$ denote the $L_1(Q)$-norm of a Borel measurable function $f: \overline{\mathbb{R}}_+^2 \to \mathbb{R}$. For $\varepsilon > 0$, the $L_1(Q)$ ε-ball around f is the set of Borel functions $\{g: \|f-g\|_Q < \varepsilon\}$. For a family of functions $\mathscr{V}$, the $(\varepsilon, L_1(Q))$-covering number $N(\varepsilon, \mathscr{V}, L_1(Q))$ is the smallest number of $L_1(Q)$ ε-balls needed to cover $\mathscr{V}$. If $\mathscr{C}$ is a VC-class, then for all $\varepsilon > 0$, the family $\mathscr{V} = \{1_C: C \in \mathscr{C}\}$ satisfies

$$\sup_{Q \in \mathscr{Q}} \log N(\varepsilon, \mathscr{V}, L_1(Q)) < \infty; \tag{5.2}$$

see Theorem 2.6.4 in van der Vaart and Wellner [28].

Recall the collection of corner sets $\mathscr{C}$ defined in §2.3:

$$\mathscr{C} = \{[x, \infty] \times [y, \infty]: x, y \in \overline{\mathbb{R}}_+\}.$$

Note that for any 3-point subset $\{x_1, x_2, x_3\} \subset \overline{\mathbb{R}}_+^2$, it is impossible for $\mathscr{C}$ to pick out all three 2-point subsets of $\{x_1, x_2, x_3\}$. Because $\mathscr{C}$ shatters no 3-point subset, it has VC-index bounded above by 3. Thus, $\mathscr{C}$ is a VC-class and $\mathscr{V} = \{1_C: C \in \mathscr{C}\}$ satisfies (5.2).

Define an envelope function for $\mathscr{V}$ as follows. Let $\pi: \overline{\mathbb{R}}_+^2 \to \overline{\mathbb{R}}_+$ be the map $\pi(x, y) = \max\{x, y\}$. Because π is continuous, (2.13) and the Skorohod representation theorem imply the existence of $\overline{\mathbb{R}}_+$-valued random variables X^r with distribution $\breve{\vartheta}^r \circ \pi^{-1}$ and X with distribution $\vartheta \circ \pi^{-1}$ such that $X^r \to X$ almost surely. Thus, there exists an $\overline{\mathbb{R}}_+$-valued random variable Y such that, almost surely,

$$Y = \sup_{r \in \mathscr{R}} X^r. \tag{5.3}$$

Let μ be the law of Y on $\overline{\mathbb{R}}_+$. Because $L_2(\mu)$ contains continuous unbounded functions, there exists a continuous, unbounded function $\psi: \overline{\mathbb{R}}_+ \to \mathbb{R}_+$ that is increasing on $[0, \infty)$, satisfies $\psi \geq 1$, and such that $\langle \psi^2, \mu\rangle < \infty$. This implies that

$$\langle (\psi \circ \pi)^2, \vartheta\rangle = \mathbf{E}[\psi(X)^2] \leq \mathbf{E}[\psi(Y)^2] < \infty. \tag{5.4}$$

Let $F = \psi \circ \pi$, and note that $1_C \leq F$ for all $C \in \mathscr{C}$. That is, F is an *envelope function* for $\mathscr{V}$. Finally, define $\overline{\mathscr{V}} = \mathscr{V} \cup \{F\}$.

Lemma 5.1. *Let $T > 0$. Then, as $r \to \infty$,*

$$\sup_{f \in \overline{\mathscr{V}}} \sup_{0 \leq s \leq t \leq T} |\langle f, \bar{\mathscr{L}}^r(s, t)\rangle - \lambda^r (t-s) \langle f, \breve{\vartheta}^r\rangle| \xrightarrow{\mathbf{P}^r} 0. \tag{5.5}$$

Proof. Let $\varepsilon > 0$. By (5.1), it suffices to show that

$$\limsup_{r \to \infty} \mathbf{P}^r\left(\sup_{f \in \overline{\mathscr{V}}} \sup_{t \in [0, T]} |\langle f, \bar{\mathscr{L}}^r(t)\rangle - \lambda^r t \langle f, \breve{\vartheta}^r\rangle| > \varepsilon\right) \leq \varepsilon.$$

Note that the above event is measurable for each r because it can be rewritten using the suprema over rational t, and $f = 1_C$ with C having rational or infinite corner coordinates x and y. Because $\langle f, \bar{\mathscr{L}}^r(t)\rangle$ and $\lambda^r t \langle f, \breve{\vartheta}^r\rangle$ are nondecreasing in t for each fixed $f \in \overline{\mathscr{V}}$, it suffices to show that for each fixed $t \in [0, T]$,

$$\limsup_{r \to \infty} \mathbf{P}^r\left(\sup_{f \in \overline{\mathscr{V}}} |\langle f, \bar{\mathscr{L}}^r(t)\rangle - \lambda^r t \langle f, \breve{\vartheta}^r\rangle| > \varepsilon\right) \leq \varepsilon.$$

Because

$$\langle f, \bar{\mathscr{L}}^r(t)\rangle - \lambda^r t \langle f, \breve{\vartheta}^r\rangle = \langle f, \breve{\vartheta}^r\rangle (\bar{E}^r(t) - \lambda^r t) + \bar{E}^r(t)\left(\frac{\langle f, \bar{\mathscr{L}}^r(t)\rangle}{\bar{E}^r(t)} - \langle f, \breve{\vartheta}^r\rangle\right)$$

(with the convention that division by zero equals zero), it suffices to show the two bounds

$$\limsup_{r\to\infty} \mathbf{P}^r\left(\sup_{f\in\bar{\mathscr{V}}} |\langle f, \breve{\vartheta}^r\rangle(\bar{E}^r(t)-\lambda^r t)| > \frac{\varepsilon}{2}\right) \le \frac{\varepsilon}{2},$$

$$\limsup_{r\to\infty} \mathbf{P}^r\left(\sup_{f\in\bar{\mathscr{V}}} \left|\bar{E}^r(t)\left(\frac{\langle f, \bar{\mathscr{L}}^r(t)\rangle}{\bar{E}^r(t)} - \langle f, \breve{\vartheta}^r\rangle\right)\right| > \frac{\varepsilon}{2}\right) \le \frac{\varepsilon}{2}. \tag{5.6}$$

The first inequality follows from assumption (2.12) and by observing that

$$\sup_{r\in\mathscr{R}}\sup_{f\in\bar{\mathscr{V}}} \langle f, \breve{\vartheta}^r\rangle \le \sup_{r\in\mathscr{R}} \langle F, \breve{\vartheta}^r\rangle = \sup_{r\in\mathscr{R}} \mathbf{E}[\psi(X^r)] \le \mathbf{E}[\psi(Y)] < \infty, \tag{5.7}$$

which follows from (5.3) and (5.4). To show (5.6), it suffices to verify three assumptions of Theorem 2.8.1 in van der Vaart and Wellner [28]. Observe that for each $n\in\mathbb{N}$ and $(e_1,\dots,e_n)\in\mathbb{R}^n$, the function

$$(x_1,\dots,x_n) \to \sup_{f\in\bar{\mathscr{V}}} \sum_{i=1}^n e_i f(x_i)$$

is measurable on the completion of $(\bar{\mathbb{R}}^2_+, \mathscr{B}, \breve{\vartheta}^r)^n$ for each $r\in\mathscr{R}$. Thus, $\bar{\mathscr{V}}$ is a $\breve{\vartheta}^r$*-measurable class* for each $r\in\mathscr{R}$; see Definition 2.3.3 in van der Vaart and Wellner [28]. Moreover, $\bar{\mathscr{V}}$ is uniformly bounded above by the envelope function F, and

$$\lim_{M\to\infty}\sup_{r\in\mathscr{R}} \langle F1_{\{F>M\}}, \breve{\vartheta}^r\rangle = 0$$

by Markov's inequality, (5.3), and (5.4). Lastly, $\bar{\mathscr{V}}$ satisfies the finite entropy bound (5.2) because $N(\varepsilon, \bar{\mathscr{V}}, L_1(Q)) \le N(\varepsilon, \mathscr{V}, L_1(Q)) + 1$ and $\mathscr{C}$ is a VC-class. The previous three observations imply that the assumptions of Theorem 2.8.1 in van der Vaart and Wellner [28] are satisfied. Consequently, $\bar{\mathscr{V}}$ is *Glivenko-Cantelli, uniformly in* r. That is, for every $\delta>0$, there exists an n_δ such that $n\ge n_\delta$ implies

$$\sup_{r\in\mathscr{R}} \mathbf{P}^r\left(\sup_{m\ge n}\sup_{f\in\bar{\mathscr{V}}}\left|\frac{1}{m}\sum_{i=1}^m f(B_i^r, D_i^r r^{-1}) - \langle f, \breve{\vartheta}^r\rangle\right| > \delta\right) \le \delta. \tag{5.8}$$

Choose $\delta = \min\{\varepsilon/2, \varepsilon/(4\lambda T)\}$. The left side of (5.6) is bounded above by

$$\limsup_{r\to\infty} \mathbf{P}^r(\bar{E}^r(t) > 2\lambda T) + \limsup_{r\to\infty} \mathbf{P}^r\left(\sup_{f\in\bar{\mathscr{V}}}\left|\frac{\langle f, \bar{\mathscr{L}}^r(t)\rangle}{\bar{E}^r(t)} - \langle f, \breve{\vartheta}^r\rangle\right| > \frac{\varepsilon}{4\lambda T}\right).$$

The first term equals zero by (2.12). For the second term, rewrite

$$\frac{\langle f, \bar{\mathscr{L}}^r(t)\rangle}{\bar{E}^r(t)} = \frac{1}{E^r(rt)} \sum_{i=1}^{E^r(rt)} f(B_i^r, D_i^r r^{-1})$$

and bound each probability in the second term by

$$\mathbf{P}^r(E^r(rt) < n_\delta) + \mathbf{P}^r\left(\left|\sup_{m\ge n_\delta}\sup_{f\in\bar{\mathscr{V}}} \frac{1}{m}\sum_{i=1}^m f(B_i^r, D_i^r r^{-1}) - \langle f, \breve{\vartheta}^r\rangle\right| > \frac{\varepsilon}{4\lambda T}\right). \tag{5.9}$$

By (2.12), the first term in (5.9) converges to zero as $r\to\infty$. By (5.8), the second term is bounded above by $\delta \le \varepsilon/2$, uniformly in $r\in\mathscr{R}$. This implies (5.6). □

5.2. Fluid scaled dynamic equation. Using (2.7), it is easy to see that the fluid scaled state descriptor of the rth model satisfies the following equation almost surely: for each Borel set $A\in\mathscr{B}$, and all $t, h\ge 0$,

$$\bar{\mathscr{Z}}^r(t+h)(A)=\bar{\mathscr{Z}}^r(t)(A+(\bar{S}^r(t,t+h),h))+\frac{1}{r}\sum_{i=r\bar{E}^r(t)+1}^{r\bar{E}^r(t+h)}1_A^+(\bar{B}_i^r(t+h),\bar{D}_i^r(t+h)). \tag{5.10}$$

Subsequent proofs use estimates obtained from this equation. Two estimates result from bounding the summands in (5.10) by one and optionally bounding the first term on the right side by its total mass; for each $A\in\mathscr{B}$ and $t, h\ge 0$,

$$\begin{aligned}\bar{\mathscr{Z}}^r(t+h)(A)&\le\bar{\mathscr{Z}}^r(t)(A+(\bar{S}^r(t,t+h),h))+\bar{\mathscr{L}}^r(t,t+h)(\overline{\mathbb{R}}_+^2)\\&\le\bar{\mathscr{Z}}^r(t)(\overline{\mathbb{R}}_+^2)+\bar{\mathscr{L}}^r(t,t+h)(\overline{\mathbb{R}}_+^2).\end{aligned}\tag{5.11}$$

Two more estimates follow from (5.10) by simply ignoring any arrivals; for each $A\in\mathscr{B}$ and $t, h\ge 0$,

$$\bar{\mathscr{Z}}^r(t)(A+(\bar{S}^r(t,t+h),h))\le\bar{\mathscr{Z}}^r(t+h)(A)\le\bar{\mathscr{Z}}^r(t+h)(\overline{\mathbb{R}}_+^2).\tag{5.12}$$

5.3. Compact containment. This section establishes the compact containment property needed to prove tightness.

LEMMA 5.2. *Let $T>0$ and $\eta>0$. There exists a compact set $\mathbf{K}\subset\mathbf{M}$ such that*

$$\liminf_{r\to\infty}\mathbf{P}^r(\bar{\mathscr{Z}}^r(t)\in\mathbf{K}\text{ for all }t\in[0,T])\ge 1-\eta.\tag{5.13}$$

PROOF. A set $\mathbf{K}\subset\mathbf{M}$ is relatively compact if $\sup_{\xi\in\mathbf{K}}\xi(\overline{\mathbb{R}}_+^2)<\infty$, and if there exists a sequence of nested compact sets $K_n\subset\overline{\mathbb{R}}_+^2$ such that $\bigcup_{n\in\mathbb{N}}K_n=\overline{\mathbb{R}}_+^2$ and

$$\lim_{n\to\infty}\sup_{\xi\in\mathbf{K}}\xi(K_n^c)=0,$$

where K_n^c denotes the complement of K_n; see Kallenberg [17, Theorem A 7.5.] Consider the nested sequence of compact sets in $\overline{\mathbb{R}}_+^2$ given by

$$K_n=([0,n]\times[0,n])\cup([0,n]\times\{\infty\})\cup(\{\infty\}\times[0,n])\cup(\{\infty\}\times\{\infty\}),\quad n\in\mathbb{N}.$$

By (2.14), $\bar{\mathscr{Z}}^r(0)\xrightarrow{\mathbf{w}}\zeta_0$ in distribution, and so the sequence $\{\bar{\mathscr{Z}}^r(0)\}$ is tight. Thus, there is a compact set $\mathbf{K}_0\subset\mathbf{M}$ such that

$$\liminf_{r\to\infty}\mathbf{P}^r(\bar{\mathscr{Z}}^r(0)\in\mathbf{K}_0)\ge 1-\frac{\eta}{2}.\tag{5.14}$$

Let $M_0=\sup_{\xi\in\mathbf{K}_0}\xi(\overline{\mathbb{R}}_+^2)$, and let $a_n=\sup_{\xi\in\mathbf{K}_0}\xi(K_n^c)$ for each $n\in\mathbb{N}$. Because $\mathbf{K}_0$ is compact, $M_0<\infty$ and there exists a sequence of nested compact sets $J_n\subset\overline{\mathbb{R}}_+^2$ such that $\bigcup_{n\in\mathbb{N}}J_n=\overline{\mathbb{R}}_+^2$ and $\lim_{n\to\infty}\sup_{\xi\in\mathbf{K}_0}\xi(J_n^c)=0$. Because $J_n\subset K_{k(n)}$ for each $n\in\mathbb{N}$ and sufficiently large $k(n)\in\mathbb{N}$, it follows that $a_n\to 0$ as $n\to\infty$.

Recall the definition from §5.1 of the envelope function $F=\psi\circ\pi$ for the family $\bar{\mathscr{V}}$. By (2.12) and (5.7), the constant $M=\sup_{r\in\mathscr{R}}(\lambda^rT\langle F,\breve{\vartheta}^r\rangle+1)$ is finite. Let $\mathbf{K}$ be the closure of the set

$$\{\xi\in\mathbf{M}\colon\ \xi(\overline{\mathbb{R}}_+^2)\le M_0+M\text{ and }\xi(K_n^c)\le a_n+\psi(n)^{-1}M\text{ for all }n\in\mathbb{N}\}.$$

Because $a_n+\psi(n)^{-1}M\to 0$ as $n\to\infty$, the set $\mathbf{K}$ is compact in $\mathbf{M}$.

For each $r\in\mathscr{R}$, denote the event in (5.14) by Ω_0^r and define the event

$$\Omega_1^r=\{\langle F,\bar{\mathscr{L}}^r(T)\rangle\le\lambda^rT\langle F,\breve{\vartheta}^r\rangle+1\}.$$

By (5.14) and Lemma 5.1, $\liminf_{r\to\infty}\mathbf{P}^r(\Omega_0^r\cap\Omega_1^r)\ge 1-\eta$. Fix $\omega\in\Omega_0^r\cap\Omega_1^r$ and $t\in[0,T]$, and assume for the remainder of the proof that all random objects are evaluated at this ω. Then, it suffices to show that $\bar{\mathscr{Z}}^r(t)\in\mathbf{K}$.

By (5.11),

$$\bar{\mathscr{Z}}^r(t)(\overline{\mathbb{R}}_+^2)\le\bar{\mathscr{Z}}^r(0)(\overline{\mathbb{R}}_+^2)+\bar{\mathscr{L}}^r(t)(\overline{\mathbb{R}}_+^2).$$

Because $\bar{\mathscr{L}}^r(t)(\overline{\mathbb{R}}_+^2)=\langle 1,\bar{\mathscr{L}}^r(t)\rangle\le\langle 1,\bar{\mathscr{L}}^r(T)\rangle\le\langle F,\bar{\mathscr{L}}^r(T)\rangle$, the definitions of Ω_0^r, Ω_1^r, and M imply that

$$\bar{\mathscr{Z}}^r(t)(\overline{\mathbb{R}}_+^2)\le M_0+M.\tag{5.15}$$

Fix $n \in \mathbb{N}$. By (5.10),

$$\bar{\mathcal{Z}}^r(t)(K_n^c) = \bar{\mathcal{Z}}^r(0)(K_n^c + (\bar{S}^r(0,t),t)) + \frac{1}{r}\sum_{i=1}^{r\bar{E}^r(t)} 1_{K_n^c}^+(\bar{B}_i^r(t), \bar{D}_i^r(t)).$$

The shape of the set K_n^c implies that

$$K_n^c + (S(0,t),t) \subset K_n^c \quad \text{and} \quad 1_{K_n^c}^+(\bar{B}_i^r(t), \bar{D}_i^r(t)) \leq 1_{K_n^c}(B_i^r, D_i^r r^{-1}),$$

for $i = 1, \ldots, r\bar{E}^r(t)$. Thus,

$$\bar{\mathcal{Z}}^r(t)(K_n^c) \leq \bar{\mathcal{Z}}^r(0)(K_n^c) + \langle 1_{K_n^c}, \bar{\mathcal{L}}^r(t)\rangle.$$

By definition of ψ, F, and by Markov's inequality, $1_{K_n^c} \leq \psi(n)^{-1}F$. So,

$$\bar{\mathcal{Z}}^r(t)(K_n^c) \leq \bar{\mathcal{Z}}^r(0)(K_n^c) + \psi(n)^{-1}\langle F, \bar{\mathcal{L}}^r(t)\rangle.$$

Because $\langle F, \bar{\mathcal{L}}^r(t)\rangle \leq \langle F, \bar{\mathcal{L}}^r(T)\rangle$, the definitions of Ω_0^r, Ω_1^r, and M imply that

$$\bar{\mathcal{Z}}^r(t)(K_n^c) \leq a_n + \psi(n)^{-1}M. \tag{5.16}$$

Equations (5.15) and (5.16) imply that $\bar{\mathcal{Z}}^r(t) \in \mathbf{K}$. $\square$

5.4. Asymptotic regularity. The second and main step necessary to prove tightness is to bound the probability that the process $\bar{\mathcal{Z}}^r(\cdot)$ oscillates. Oscillations may result from sudden arrivals or departures of a large amount of mass. Sudden arrivals are controlled by the regularity of the arrival process. To show that sudden departures are unlikely as well, we show that $\bar{\mathcal{Z}}^r(\cdot)$ assigns arbitrarily small mass to the boundaries of the sets $C \in \mathscr{C}$. This is phrased in terms of κ-enlargements of the boundaries of these sets (forming a collection of L-shaped sets). For $C \in \mathscr{C}$ and $\kappa > 0$, let ∂_C denote the boundary of C in $\bar{\mathbb{R}}_+^2$ and let

$$\partial_C^\kappa = \left\{ w \in \bar{\mathbb{R}}_+^2 \colon \inf_{z \in \partial_C} \|w - z\| < \kappa \right\}$$

be the κ-enlargement in $\bar{\mathbb{R}}_+^2$ of its boundary, where the infimum over the empty set equals ∞. (Note that ∂_C and, therefore, also ∂_C^κ is empty for the corner set $\bar{\mathbb{R}}_+^2$. Note also that $\partial_C^\kappa = ((x-\kappa)^+, x+\kappa) \times \{\infty\}$ for a corner set of the form $[x,\infty] \times \{\infty\}$ with $x \in [0,\infty)$.) The following lemma establishes the result for the initial condition $\bar{\mathcal{Z}}^r(0)$.

Lemma 5.3. *For all $\varepsilon, \eta > 0$, there exists a $\kappa > 0$ such that*

$$\liminf_{r\to\infty} \mathbf{P}^r\left(\sup_{C\in\mathscr{C}} \bar{\mathcal{Z}}^r(0)(\partial_C^\kappa) \leq \varepsilon\right) \geq 1 - \eta. \tag{5.17}$$

Proof. Fix $\varepsilon, \eta > 0$ and let $\bar{\mathcal{Z}}_1^r(0)(\cdot) = \bar{\mathcal{Z}}^r(0)(\cdot \times \bar{\mathbb{R}}_+)$ and $\bar{\mathcal{Z}}_2^r(0)(\cdot) = \bar{\mathcal{Z}}^r(0)(\bar{\mathbb{R}}_+ \times \cdot)$. For each $C \in \mathscr{C}$ and $\kappa > 0$,

$$\partial_C^\kappa \subset ([x, x+2\kappa] \times \bar{\mathbb{R}}_+) \cup (\bar{\mathbb{R}}_+ \times [y, y+2\kappa])$$

for some $(x,y) \in \mathbb{R}_+^2 = [0,\infty) \times [0,\infty)$. Thus, it suffices to show that for $i = 1, 2$, there exists a $\kappa > 0$ such that

$$\liminf_{r\to\infty} \mathbf{P}^r\left(\sup_{x\in[0,\infty)} \bar{\mathcal{Z}}_i^r(0)([x, x+2\kappa]) \leq \frac{\varepsilon}{2}\right) \geq 1 - \frac{\eta}{2}. \tag{5.18}$$

We prove the statement for $i = 1$; the proof is identical for $i = 2$.

The projection $(x,y) \mapsto x$ is continuous, so (2.14) implies that $\bar{\mathcal{Z}}_1^r(0)$ converges in distribution to $\zeta_0(\cdot \times \bar{\mathbb{R}}_+)$ as $r \to \infty$. Because $\zeta_0(\cdot \times \bar{\mathbb{R}}_+)$ is free of atoms in $[0,\infty)$, there exists a $\kappa > 0$ such that

$$\sup_{x\in[0,\infty)} \zeta_0([x, x+4\kappa] \times \bar{\mathbb{R}}_+) \leq \frac{\varepsilon}{2}. \tag{5.19}$$

(If (5.19) fails, it is easy to construct an atom of $\zeta_0(\cdot \times \bar{\mathbb{R}}_+)$.) Moreover, there exists a constant M such that

$$\zeta_0([M,\infty) \times \bar{\mathbb{R}}_+) \leq \frac{\varepsilon}{2}. \tag{5.20}$$

Let $N = \lceil M/\kappa \rceil + 1$, where $\lceil x \rceil$ denotes the smallest integer $n \geq x$. For $n = 0, \ldots, N-1$, define the set $I_n = [n\kappa, (n+4)\kappa]$ and define $I_N = [M, \infty)$. Note that for every $x \in [0, \infty)$, there is an $n \leq N$ such that $[x, x+2\kappa] \subset I_n$. To prove (5.18), it therefore suffices to show that

$$\liminf_{r\to\infty} \mathbf{P}^r\left(\max_{n\leq N} \bar{\mathcal{Z}}_1^r(0)(I_n) \leq \frac{\varepsilon}{2}\right) \geq 1 - \frac{\eta}{2}. \tag{5.21}$$

Let $\mathbf{M}(\overline{\mathbb{R}}_+)$ denote the space of finite nonnegative Borel measures on $\overline{\mathbb{R}}_+$, endowed with the weak topology. Let $\mathbf{A} = \{\xi \in \mathbf{M}(\overline{\mathbb{R}}_+)\colon \max_{n\leq N} \xi(I_n) < \varepsilon/2\}$, and suppose that a sequence $\{\xi_k\} \subset \mathbf{M}(\overline{\mathbb{R}}_+)$ satisfies $\xi_k \xrightarrow{\mathbf{w}} \xi$ for some $\xi \in \mathbf{A}$. Because the sets I_n are closed, the Portmanteau theorem (adapted to finite measures) implies that

$$\limsup_{k\to\infty} \xi_k(I_n) \leq \xi(I_n) < \frac{\varepsilon}{2} \quad \text{for all } n \leq N.$$

Hence, $\xi_k \in \mathbf{A}$ for sufficiently large k, which implies that $\mathbf{A}$ is open in $\mathbf{M}(\overline{\mathbb{R}}_+)$. Thus, a second application of the Portmanteau theorem yields

$$\liminf_{r\to\infty} \mathbf{P}^r(\bar{\mathcal{Z}}_1^r(0) \in \mathbf{A}) \geq \mathbf{P}(\zeta_0(\cdot \times \overline{\mathbb{R}}_+) \in \mathbf{A}) = 1,$$

which implies (5.21). □

The regularity result is now shown for the entire state descriptor $\bar{\mathcal{Z}}^r(\cdot)$.

Lemma 5.4. *Let $T > 0$ and $\varepsilon, \eta > 0$. There exists a $\kappa > 0$ such that*

$$\liminf_{r\to\infty} \mathbf{P}^r\left(\sup_{C\in\mathscr{C}} \sup_{t\in[0,T]} \bar{\mathcal{Z}}^r(t)(\partial_C^\kappa) \leq \varepsilon\right) \geq 1 - \eta. \tag{5.22}$$

Proof. By Lemmas 5.1, 5.2, and 5.3, there exists a compact $\mathbf{K} \subset \mathbf{M}$ and a $\kappa_0 > 0$, such that for all $\delta > 0$, the events

$$\Omega_1^r = \left\{\sup_{C\in\mathscr{C}} \bar{\mathcal{Z}}^r(0)(\partial_C^{\kappa_0}) \leq \frac{\varepsilon}{2}\right\},$$

$$\Omega_2^r = \left\{\sup_{C\in\mathscr{C}} \sup_{0\leq s\leq t\leq T} |\bar{\mathcal{L}}^r(s,t)(C) - \lambda^r(t-s)\check{\vartheta}^r(C)| \leq \delta\right\},$$

$$\Omega_3^r = \left\{\bar{\mathcal{Z}}^r(t) \in \mathbf{K} \text{ for all } t \in [0,T]\right\},$$

$$\Omega_0^r = \Omega_1^r \cap \Omega_2^r \cap \Omega_3^r,$$

satisfy

$$\liminf_{r\to\infty} \mathbf{P}^r(\Omega_0^r) \geq 1 - \eta. \tag{5.23}$$

Recall the compact sets K_n defined in the proof of Lemma 5.2. Because $\mathbf{K}$ is compact, there exists a finite $M \geq 1$ and an integer $R < \infty$ such that

$$\sup_{\xi\in\mathbf{K}} \xi(\mathbb{R}_+^2) \leq M, \tag{5.24}$$

$$\sup_{\xi\in\mathbf{K}} \xi(K_R^c) \leq \frac{\varepsilon}{2}. \tag{5.25}$$

Let $\lambda^* = \sup_{r\in\mathscr{R}} \lambda^r$, which is finite by (2.12). Fix

$$h = \varepsilon(8\lambda^*)^{-1}, \qquad \kappa = \min\{\kappa_0, h(2M)^{-1}\}, \qquad \text{and} \qquad \delta = \varepsilon \min\{(8\lceil RMh^{-1}\rceil)^{-1}, 2^{-1}\}.$$

For $r \in \mathscr{R}$, let Ω_*^r denote the event in (5.22). By (5.23), it suffices to show that $\Omega_0^r \subset \Omega_*^r$. Let $\omega \in \Omega_0^r$ be arbitrary; for the remainder of the proof, all random objects are evaluated at this ω.

Consider any $r \in \mathscr{R}$, $t \in [0,T]$, and $C \in \mathscr{C}$. We must show that $\bar{\mathcal{Z}}^r(t)(\partial_C^\kappa) \leq \varepsilon$. Define the random time

$$\tau_1 = \sup\{s \leq t\colon \langle 1, \bar{\mathcal{Z}}^r(s)\rangle = 0\}$$

if the supremum exists, and define $\tau_1 = 0$ otherwise. Let $\tau = \max\{\tau_1, t - RM\}$. We first show that

$$\bar{\mathcal{Z}}^r(\tau)(\partial_C^\kappa + (\bar{S}^r(\tau,t), t-\tau)) \leq \frac{\varepsilon}{2}. \tag{5.26}$$

If $\tau = 0$, this follows from the definition of Ω_1^r because $\kappa \le \kappa_0$, because

$$\partial_C^\kappa + (\bar{S}^r(\tau, t), t-\tau) \subset \partial^\kappa_{C+(\bar{S}^r(\tau,t), t-\tau)},$$

and because $\mathscr{C}$ is closed under positive translation. Suppose $\tau = \tau_1 > 0$. Then, there is a sequence $\{\tau_n\}$, with $\tau_n \uparrow \tau$, such that $\langle 1, \bar{\mathscr{Z}}^r(\tau_n)\rangle = 0$ for all n. In this case, (5.11) and the definition of Ω_2^r imply that, for all n,

$$\bar{\mathscr{Z}}^r(\tau)(\partial_C^\kappa + (\bar{S}^r(\tau,t), t-\tau)) \le \bar{\mathscr{Z}}^r(\tau_n)(\bar{\mathbb{R}}_2^+) + \bar{\mathscr{L}}^r(\tau_n, \tau)(\bar{\mathbb{R}}_2^+) \le \lambda^r(\tau-\tau_n) + \delta.$$

Letting $\tau_n \uparrow \tau$ yields

$$\bar{\mathscr{Z}}^r(\tau)(\partial_C^\kappa + (\bar{S}^r(\tau,t), t-\tau)) \le \delta \le \frac{\varepsilon}{2}.$$

Suppose that $\tau = t - RM$. Because $\langle 1, \bar{\mathscr{Z}}^r(s)\rangle > 0$ for all $s \in (\tau, t]$, the definition of Ω_3^r and (5.24) imply that

$$\bar{S}^r(\tau, t) = \int_{t-RM}^t \langle 1, \bar{\mathscr{Z}}^r(s)\rangle^{-1}\, ds \ge R.$$

Thus, by the definition of Ω_3^r and (5.25),

$$\bar{\mathscr{Z}}^r(\tau)(\partial_C^\kappa + (\bar{S}^r(\tau,t), t-\tau)) \le \bar{\mathscr{Z}}^r(\tau)(K_R^c) \le \frac{\varepsilon}{2},$$

which proves (5.26).

By (5.10),

$$\bar{\mathscr{Z}}^r(t)(\partial_C^\kappa) = \bar{\mathscr{Z}}^r(\tau)(\partial_C^\kappa + (\bar{S}^r(\tau,t), t-\tau)) + \frac{1}{r}\sum_{i=r\bar{E}^r(\tau)+1}^{r\bar{E}^r(t)} 1^+_{\partial_C^\kappa}(\bar{B}_i^r(t), \bar{D}_i^r(t)). \tag{5.27}$$

Let I denote the second right-hand term in (5.27). By (5.26), it remains to show that $I \le \varepsilon/2$. Let $N = \lceil (t-\tau)h^{-1}\rceil$ and, for each $n = 0, \ldots, N-1$, let $t_n = \tau + nh$ and $t^n = \min\{t_{n+1}, t\}$. Then, using the inequality $1^+_{\partial_C^\kappa}(\cdot,\cdot) \le 1_{\partial_C^\kappa}(\cdot,\cdot)$,

$$I \le \sum_{n=0}^{N-1} \frac{1}{r} \sum_{i=r\bar{E}^r(t_n)+1}^{r\bar{E}^r(t^n)} 1_{\partial_C^\kappa}(\bar{B}_i^r(t), \bar{D}_i^r(t)). \tag{5.28}$$

Consider $n \in \{0, \ldots, N-1\}$ and i such that $U_i^r r^{-1} \in (t_n, t^n]$. Observe that

$$\bar{S}^r(t^n, t) \le \bar{S}^r(U_i^r r^{-1}, t) \le \bar{S}^r(t_n, t). \tag{5.29}$$

By definition,

$$1_{\partial_C^\kappa}(\bar{B}_i^r(t), \bar{D}_i^r(t)) = 1_{\partial_C^\kappa + (\bar{S}^r(U_i^r r^{-1}, t), t - U_i^r r^{-1})}(B_i^r, D_i^r r^{-1}). \tag{5.30}$$

So, letting

$$C_n^- = C + (\bar{S}^r(t^n, t) - \kappa, t - t^n - \kappa) \cap \bar{\mathbb{R}}_2^+,$$
$$C_n^+ = C + (\bar{S}^r(t_n, t) + \kappa, t - t_n + \kappa) \cap \bar{\mathbb{R}}_2^+,$$
$$C_n = C_n^- \backslash C_n^+,$$

it follows from (5.29) and (5.30) that

$$1_{\partial_C^\kappa}(\bar{B}_i^r(t), \bar{D}_i^r(t)) \le 1_{C_n}(B_i^r, D_i^r r^{-1}). \tag{5.31}$$

Conclude from (5.28) and (5.31) that

$$I \le \sum_{n=0}^{N-1} \frac{1}{r} \sum_{i=r\bar{E}^r(t_n)+1}^{r\bar{E}^r(t^n)} 1_{C_n}(B_i^r, D_i^r r^{-1}) = \sum_{n=0}^{N-1} (\bar{\mathscr{L}}^r(t_n, t^n)(C_n^-) - \bar{\mathscr{L}}^r(t_n, t^n)(C_n^+)).$$

For all $n < N$, $C_n^-, C_n^+ \in \mathscr{C}$ and $t^n - t_n \le h$. So, the definition of Ω_2^r implies that

$$I \le \sum_{n=0}^{N-1} (\lambda^r h \breve{\vartheta}^r(C_n) + 2\delta).$$

By definition of N, and because $t-\tau \le RM$,

$$I \le \lambda^* h \sum_{n=0}^{N-1} \breve{\vartheta}^r(C_n) + \lceil RMh^{-1} \rceil 2\delta.$$

This implies, by choice of δ, that

$$I \le \lambda^* h \sum_{n=0}^{N-1} \breve{\vartheta}^r(C_n) + \frac{\varepsilon}{4}. \tag{5.32}$$

If $n \in \{0, \ldots, N-3\}$, then

$$\bar{S}^r(t_{n+1}, t_{n+2}) \ge hM^{-1} \ge 2\kappa,$$

because $0 < \langle 1, \bar{\mathscr{Z}}^r(s)\rangle \le M$ for all $s \in (\tau, t]$ and because $h \ge \kappa 2M$ by definition of κ. Thus, for all $n \in \{0, \ldots, N-3\}$,

$$\bar{S}^r(t^n, t) - \kappa = \bar{S}^r(t_{n+1}, t_{n+2}) + \bar{S}^r(t_{n+2}, t) - \kappa \ge \bar{S}^r(t_{n+2}, t) + \kappa.$$

Hence, $C_n^- \subset C_{n+2}^+$ for all $n \in \{0, \ldots, N-3\}$ and, consequently, $C_n \cap C_{n+2} = \varnothing$. Thus, because $\breve{\vartheta}^r$ is a probability measure,

$$\sum_{n=0}^{\lfloor (N-1)/2 \rfloor} \breve{\vartheta}^r(C_{2n}) \qquad \text{and} \qquad \sum_{n=0}^{\lfloor (N-2)/2 \rfloor} \breve{\vartheta}^r(C_{2n+1})$$

are both bounded above by one. Conclude from (5.32) that

$$I \le 2\lambda^* h + \frac{\varepsilon}{4},$$

which implies, by choice of h, that $I \le \varepsilon/2$. □

5.5. Oscillation bound. This section establishes the second main ingredient for proving tightness of the state descriptors. As a metric on $\mathbf{M}$, we use the Prohorov metric (adapted to finite measures). For $\mu, \nu \in \mathbf{M}$, define

$$\mathbf{d}[\mu, \nu] = \inf\{\varepsilon > 0\colon \mu(A) \le \nu(A^\varepsilon) + \varepsilon \text{ and } \nu(A) \le \mu(A^\varepsilon) + \varepsilon \text{ for all closed } A \in \mathscr{B}\}.$$

Recall that $A^\varepsilon = \{w \in \overline{\mathbb{R}}_+^2\colon \inf_{z \in A} \|z - w\| < \varepsilon\}$ and that $\mathscr{B}$ denotes the Borel subsets of $\overline{\mathbb{R}}_+^2$.

Definition 5.5. For each $\zeta(\cdot) \in \mathbf{D}([0, \infty), \mathbf{M})$ and each $T > \delta > 0$, define the modulus of continuity on $[0, T]$ by

$$\mathbf{w}_T(\zeta(\cdot), \delta) = \sup_{t \in [0, T-\delta]} \sup_{h \in [0, \delta]} \mathbf{d}[\zeta(t+h), \zeta(t)].$$

Lemma 5.6. *For all $T > 0$ and $\varepsilon, \eta \in (0, 1)$, there exists $\delta \in (0, T)$ such that*

$$\liminf_{r \to \infty} \mathbf{P}^r(\mathbf{w}_T(\bar{\mathscr{Z}}^r(\cdot), \delta) \le \varepsilon) \ge 1 - \eta. \tag{5.33}$$

Proof. Fix $T > 0$ and $\varepsilon, \eta \in (0, 1)$, and let $\lambda^* = \sup_{r \in \mathscr{R}} \lambda^r$. For each $\kappa > 0$, define

$$L_\kappa = ([0, \kappa] \times \overline{\mathbb{R}}_+) \cup (\overline{\mathbb{R}}_+ \times [0, \kappa]).$$

By Lemmas 5.1 and 5.4, there exists $\kappa \in (0, 1)$ such that for all $\delta \in (0, T)$, the events

$$\Omega_1^r = \left\{ \sup_{t \in [0, T]} \bar{\mathscr{Z}}^r(t)(L_\kappa) \le \frac{\varepsilon}{4} \right\},$$

$$\Omega_2^r = \left\{ \sup_{t \in [0, T-\delta]} \bar{\mathscr{L}}^r(t, t+\delta)(\overline{\mathbb{R}}_+^2) \le 2\lambda^* \delta \right\},$$

$$\Omega_0^r = \Omega_1^r \cap \Omega_2^r,$$

satisfy

$$\liminf_{r \to \infty} \mathbf{P}^r(\Omega_0^r) \ge 1 - \eta. \tag{5.34}$$

Fix $\delta = \kappa\varepsilon^2(8\max\{\lambda^*, 1\})^{-1}$ and let Ω^r_* be the event in (5.33). By (5.34), it suffices to show that $\Omega^r_0 \subset \Omega^r_*$ for each r. Fix $r \in \mathcal{R}$ and $\omega \in \Omega^r_0$; for the remainder of the proof all random objects are evaluated at this ω. Fix $t \in [0, T-\delta]$, $h \in [0, \delta]$ and let $A \in \mathcal{B}$ be closed. It suffices to show the two inequalities,

$$\bar{\mathcal{Z}}^r(t)(A) \le \bar{\mathcal{Z}}^r(t+h)(A^\varepsilon) + \varepsilon \quad \text{and} \tag{5.35}$$

$$\bar{\mathcal{Z}}^r(t+h)(A) \le \bar{\mathcal{Z}}^r(t)(A^\varepsilon) + \varepsilon. \tag{5.36}$$

To show (5.35), use the definition of Ω^r_1 to write

$$\begin{aligned}\bar{\mathcal{Z}}^r(t)(A) &\le \bar{\mathcal{Z}}^r(t)(L_\kappa) + \bar{\mathcal{Z}}^r(t)(A \cap L^c_\kappa)\\ &\le \frac{\varepsilon}{4} + \bar{\mathcal{Z}}^r(t)(A\cap L^c_\kappa).\end{aligned} \tag{5.37}$$

Let $I = \{s \in [t, t+h] : \langle 1, \bar{\mathcal{Z}}^r(s)\rangle < \varepsilon/4\}$. Suppose $I = \varnothing$. Then, $\langle 1, \bar{\mathcal{Z}}^r(s)\rangle \ge \varepsilon/4$ for all $s \in [t, t+h]$, which implies that

$$\|(\bar{S}^r(t, t+h), h)\| \le \int_t^{t+\delta} \langle 1, \bar{\mathcal{Z}}^r(s)\rangle^{-1}\,\mathrm{d}s + \delta \le \frac{4\delta}{\varepsilon} + \delta < \min\{\varepsilon, \kappa\}. \tag{5.38}$$

Consequently, $(x, y) \in A \cap L^c_\kappa$ implies $(x, y) - (\bar{S}^r(t, t+h), h) \in A^\varepsilon$, and so

$$A \cap L^c_\kappa \subset A^\varepsilon + (\bar{S}^r(t, t+h), h). \tag{5.39}$$

Deduce from (5.37) that

$$\bar{\mathcal{Z}}^r(t)(A) \le \frac{\varepsilon}{4} + \bar{\mathcal{Z}}^r(t)(A^\varepsilon + (\bar{S}^r(t, t+h), h)).$$

Apply (5.12) to get

$$\bar{\mathcal{Z}}^r(t)(A) \le \frac{\varepsilon}{4} + \bar{\mathcal{Z}}^r(t+h)(A^\varepsilon). \tag{5.40}$$

Suppose $I \ne \varnothing$ and let $\tau = \inf I$. Then, $\langle 1, \bar{\mathcal{Z}}^r(\tau)\rangle \le \varepsilon/4$ by right continuity. Because $\langle 1, \bar{\mathcal{Z}}^r(s)\rangle \ge \varepsilon/4$ for all $s \in [t, \tau)$,

$$\|(\bar{S}^r(t, \tau), \tau - t)\| \le \int_t^{\tau} \langle 1, \bar{\mathcal{Z}}^r(s)\rangle^{-1}\,\mathrm{d}s + \delta \le \frac{4\delta}{\varepsilon} + \delta < \kappa. \tag{5.41}$$

By (5.37) and (5.41),

$$\bar{\mathcal{Z}}^r(t)(A) \le \frac{\varepsilon}{4} + \bar{\mathcal{Z}}^r(t)(L^c_\kappa) \le \frac{\varepsilon}{4} + \bar{\mathcal{Z}}^r(t)(\mathbb{R}^2_+ + (\bar{S}^r(t, \tau), \tau - t)).$$

Apply (5.12) to get

$$\bar{\mathcal{Z}}^r(t)(A) \le \frac{\varepsilon}{4} + \bar{\mathcal{Z}}^r(\tau)(\mathbb{R}^2_+) \le \frac{\varepsilon}{2}. \tag{5.42}$$

So, (5.35) follows because either (5.40) or (5.42) holds.

To show (5.36), use (5.11) and the definitions of Ω^r_2 and δ to obtain

$$\begin{aligned}\bar{\mathcal{Z}}^r(t+h)(A) &\le \bar{\mathcal{Z}}^r(t)(A + (\bar{S}^r(t, t+h), h)) + \bar{\mathcal{L}}^r(t, t+h)(\mathbb{R}^2_+)\\ &\le \bar{\mathcal{Z}}^r(t)(A + (\bar{S}^r(t, t+h), h)) + \frac{\varepsilon}{4}.\end{aligned} \tag{5.43}$$

If $I = \varnothing$, then (5.38) implies that $A + (\bar{S}^r(t, t+h), h) \subset A^\varepsilon$. So, (5.43) yields

$$\bar{\mathcal{Z}}^r(t+h)(A) \le \bar{\mathcal{Z}}^r(t)(A^\varepsilon) + \frac{\varepsilon}{4}.$$

If $I \ne \varnothing$, then by (5.11), the definition of Ω^r_2, and the choice of δ,

$$\bar{\mathcal{Z}}^r(t+h)(A) \le \bar{\mathcal{Z}}^r(\tau)(\mathbb{R}^2_+) + \bar{\mathcal{L}}^r(\tau, t+h)(\mathbb{R}^2_+) \le \frac{\varepsilon}{4} + 2\lambda^*\delta \le \frac{\varepsilon}{2}.$$

In both cases, (5.36) holds. Conclude from (5.35) and (5.36) that

$$\mathbf{d}[\bar{\mathcal{Z}}^r(t), \bar{\mathcal{Z}}^r(t+h)] \le \varepsilon.$$

Because $t \in [0, T-\delta]$ and $h \in [0, \delta]$ were arbitrary,

$$\mathbf{w}_T(\bar{\mathcal{Z}}^r(\cdot), \delta) \le \varepsilon,$$

which implies that $\omega \in \Omega^r_*$. $\square$

6. Limiting fluid equations. This section contains the proof of Theorem 2.3. Tightness of the sequence $\{\bar{\mathscr{Z}}^r(\cdot)\}$ follows immediately from Lemmas 5.2 and 5.6. Because $\{\bar{\mathscr{Z}}^r(\cdot)\}$ is tight, there exists a subsequence $\{q\} \subset \mathscr{R}$ and a process $\mathscr{Z}(\cdot)$ in $\mathbf{D}([0,\infty),\mathbf{M})$ such that $\bar{\mathscr{Z}}^q(\cdot) \Rightarrow \mathscr{Z}(\cdot)$ as $q \to \infty$. We must show that $\mathscr{Z}(\cdot)$ is almost surely a measure-valued fluid model solution for the data $(\lambda, \vartheta, \zeta_0)$. This is accomplished in Lemmas 6.1 and 6.2 and Theorem 6.3 below. Finally, if (2.11) holds, then a measure-valued fluid model solution for $(\lambda, \vartheta, \zeta_0)$ is unique by Theorem 2.2. In this case, the law of the limit point $\mathscr{Z}(\cdot)$ is unique and so $\bar{\mathscr{Z}}^r(\cdot) \Rightarrow \mathscr{Z}(\cdot)$ as $r \to \infty$.

Let $Z(\cdot) = \langle 1, \mathscr{Z}(\cdot)\rangle$ be the total mass process for $\mathscr{Z}(\cdot)$, and let $S(u,v) = \int_u^v 1/Z(s)\,ds$ for all $v \ge u \ge 0$. To show that $\mathscr{Z}(\cdot)$ is almost surely a measure-valued fluid model solution, note first that $\mathscr{Z}(\cdot)$ is almost surely continuous by Lemma 5.6. Note also that, by (2.14), $\mathscr{Z}(0) = \zeta_0$ almost surely. It remains to show that properties (i) and (ii) of Definition 2.1 are satisfied almost surely by $\mathscr{Z}(\cdot)$. The next result establishes (i).

Lemma 6.1. *Almost surely, for all $a > 0$,*

$$\inf_{t>a} Z(t) > 0. \tag{6.1}$$

Proof. Suppose first that $\mathbf{P}(B=\infty) > 0$, and let $T > a > 0$ be arbitrary. It suffices to show that $\inf_{t\in(a,T)} Z(t) > 0$ almost surely. Assume without loss of generality that a is sufficiently small that $\mathbf{P}(B=\infty;\, D \ge a) > 0$, and that the distribution of D is continuous at a. Define the corner set $C_a = \{\infty\} \times [a,\infty]$ and let $k_a = \lambda a \vartheta(C_a)/2 > 0$. For each $q \in \mathscr{R}$ and $t \in (a,T)$,

$$\bar{Z}^q(t) = \bar{\mathscr{Z}}^q(t)(\overline{\mathbb{R}}_+^2) \ge \bar{\mathscr{Z}}^q(t)(\{\infty\}\times[0,\infty]).$$

Applying (5.10) and then dropping the first right-hand term yields

$$\bar{Z}^q(t) \ge \frac{1}{q} \sum_{i=q\bar{E}^q(t-a)+1}^{q\bar{E}^q(t)} 1^+_{\{\infty\}\times[0,\infty]}\big(\bar{B}_i^q(t), \bar{D}_i^q(t)\big). \tag{6.2}$$

Note that for $i > q\bar{E}^q(t-a)$, we have $U_i^q \ge q(t-a)$ and so $\bar{D}_i^q(t) = D_i^q q^{-1} - t + U_i^q q^{-1} \ge D_i^q q^{-1} - a$. This implies that, for each such i, $1^+_{\{\infty\}\times[0,\infty]}(\bar{B}_i^q(t), \bar{D}_i^q(t)) \ge 1_{C_a}(B_i^q, D_i^q q^{-1})$. Deduce from (6.2) that

$$\bar{Z}^q(t) \ge \bar{\mathscr{L}}^q(t-a,t)(C_a). \tag{6.3}$$

Because C_a is a ϑ-continuity set, $\lambda^q a \check{\vartheta}^q(C_a) \to \lambda a \vartheta(C_a)$ by (2.12) and (2.13). So, by (6.3), Lemma 5.1, and the definition of k_a,

$$\liminf_{q\to\infty} \mathbf{P}^q\Big(\inf_{t\in(a,T)} \bar{Z}^q(t) \ge k_a\Big) \ge \liminf_{q\to\infty} \mathbf{P}^q\Big(\inf_{t\in(a,T)} \bar{\mathscr{L}}^q(t-a,t)(C_a) \ge k_a\Big) = 1. \tag{6.4}$$

Note that $\bar{\mathscr{Z}}^q(\cdot) \Rightarrow \mathscr{Z}(\cdot)$ implies $\bar{Z}^q(\cdot) \Rightarrow Z(\cdot)$, and that the set

$$\Big\{z(\cdot) \in \mathbf{D}([0,\infty),\mathbb{R}_+)\colon \inf_{t\in(a,T)} z(t) \ge k_a\Big\}$$

is closed in the Skorohod J_1-topology. Thus, the Portmanteau theorem and (6.4) imply that

$$\mathbf{P}\Big(\inf_{t\in(a,T)} Z(t) \ge k_a\Big) \ge \liminf_{q\to\infty} \mathbf{P}^q\Big(\inf_{t\in(a,T)} \bar{Z}^q(t) \ge k_a\Big) = 1.$$

It remains to consider the case $\mathbf{P}(B=\infty)=0$. Again, we will construct a lower bound on the queue length process. The main difference with the case above is that we exploit results for overloaded PS queues without impatience developed in Puha et al. [24]. Because the idea of the lower bound is similar as before, we restrict to giving an outline. If $\mathbf{P}(B=\infty)=0$, there exists $m<\infty$ such that $\lambda\mathbf{E}[B1_{\{B\le m\}}] > 1$, and such that the distribution of B is continuous at m. Fix $a>0$, and assume without loss of generality that a is sufficiently small that $\lambda\mathbf{E}[B1_{\{B\le m;\, D>a\}}] > 1$, and that the distribution of D is continuous at a. Then, by (2.13),

$$\lim_{q\to\infty} \lambda^q\big[B_1^q 1_{\{B_1^q \le m;\, D_1^q q^{-1} > a\}}\big] = \lambda\mathbf{E}[B1_{\{B\le m;\, D>a\}}].$$

Compare $Z^q(\cdot)$ with the queue length process $\acute{Z}^q(\cdot)$ of an ordinary PS queue having arrival rate $\lambda^q_{a,m} = \lambda^q\mathbf{P}(B_1^q \le m;\, D_1^q > qa)$ and service times $B^q_{i,a,m}$, which are distributed as $B_i^q \mid B_i^q \le m;\, D_i^q > qa$. This process can be constructed from the primitives $E^q(\cdot)$, $\{B_i^q, D_i^q\}$ in an obvious way, enabling a sample-path comparison of the two processes. Assume that $\acute{Z}^q(q(t-a)) = 0$.

Observe that the number of arrivals in the ordinary PS queue during the time interval $[q(t-a), qt]$ is less than or equal to the number of arrivals during that time interval in the PS queue with impatience. Furthermore, if a job that arrived in the original system after time $q(t-a)$ departs before time qt, then this must also be the case in the ordinary PS queue, because that system had a service rate which was at least as large as in the original system. These considerations imply that $Z^q(qt) \geq \acute{Z}^q(qt)$. The ordinary queue is still overloaded (for sufficiently large q), no customer departs because of impatience, and the modified arrival process is still a renewal process. Thus, the evolution of the modified system during $[q(t-a), qt]$ has the same law as that of an overloaded $GI/GI/1$ PS queue starting at 0, in the time interval $[0, qa]$.

Because the service times in our modified system are bounded, the means converge as $q \to \infty$. The assumptions in Puha et al. [24] are therefore satisfied, and it follows that there exists a constant $k_a > 0$ such that $\lim_{q\to\infty} \acute{Z}^q(qt)/q = k_a$ almost surely. Consequently, we have $\liminf_{q\to\infty} \bar{Z}^q(t) \geq k_a$ almost surely, which implies the assertion. $\square$

Before establishing property (ii) of Definition 2.1, the following result is needed.

LEMMA 6.2. *Almost surely, for all $C \in \mathscr{C}$ and $t \geq 0$,*

$$\mathscr{Z}(t)(\partial_C) = 0. \tag{6.5}$$

PROOF. Let $T > 0$. It suffices to show the statement for all $t \in [0, T)$. Let $\{\eta_n\} \subset (0, 1)$ be a sequence such that $\sum_{n=1}^{\infty} \eta_n < \infty$. By Lemma 5.4, there exists a null sequence of positive reals $\{\kappa_n\}$ such that, for each fixed n,

$$\liminf_{q\to\infty} \mathbf{P}^q\left(\sup_{t\in[0,T]} \sup_{C\in\mathscr{C}} \bar{\mathscr{Z}}^q(t)(\partial_C^{\kappa_n}) \leq \frac{1}{n}\right) \geq 1 - \eta_n. \tag{6.6}$$

For each $n \in \mathbb{N}$, let $\mathbf{M}_n = \{\xi \in \mathbf{M}\colon \sup_{C\in\mathscr{C}} \xi(\partial_C^{\kappa_n}) \leq 1/n\}$. If a sequence $\{\xi_i\} \subset \mathbf{M}_n$ converges weakly to ξ, then for each open set $\partial_C^{\kappa_n}$ the Portmanteau theorem yields

$$\xi(\partial_C^{\kappa_n}) \leq \limsup_{i\to\infty} \xi_i(\partial_C^{\kappa_n}) \leq \frac{1}{n}.$$

Thus, $\xi \in \mathbf{M}_n$ and $\mathbf{M}_n$ is closed. By definition of the Skorohod J_1-topology, the set $\mathbf{D}_n^T = \{\zeta(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M})\colon \zeta(t) \in \mathbf{M}_n$ for all $t \in [0, T)\}$ is also closed. Apply the Portmanteau theorem and (6.6) to obtain

$$\mathbf{P}(\mathscr{Z}(\cdot) \in \mathbf{D}_n^T) \geq \liminf_{q\to\infty} \mathbf{P}^q(\bar{\mathscr{Z}}^q(\cdot) \in \mathbf{D}_n^T) \geq 1 - \eta_n.$$

By the Borel-Cantelli lemma,

$$\mathbf{P}\left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \{\mathscr{Z}(\cdot) \in \mathbf{D}_n^T\}\right) = 1.$$

Thus, there exists a finite random variable N such that almost surely,

$$\sup_{t\in[0,T)} \sup_{C\in\mathscr{C}} \mathscr{Z}(t)(\partial_C^{\kappa_n}) \leq \frac{1}{n}, \quad \text{for all } n > N. \tag{6.7}$$

Because $\partial_C \subset \partial_C^{\kappa_n}$ for all $C \in \mathscr{C}$ and $n \in \mathbb{N}$, conclude that almost surely,

$$\sup_{t\in[0,T)} \sup_{C\in\mathscr{C}} \mathscr{Z}(t)(\partial_C) = 0. \qquad \square$$

We now establish property (ii). Recall that $Z(t) = \langle 1, \mathscr{Z}(t)\rangle$ for all $t \geq 0$, and $S(u, v) = \int_u^v 1/Z(s)\,\mathrm{d}s$ for all $v \geq u \geq 0$.

THEOREM 6.3. *Almost surely, the process $\mathscr{Z}(\cdot)$ satisfies*

$$\mathscr{Z}(t)(A) = \mathscr{Z}(0)(A + (S(0,t), t)) + \lambda \int_0^t \vartheta(A + (S(s,t), t-s))\,\mathrm{d}s \tag{6.8}$$

for all $t \geq 0$ and $A \in \mathscr{B}$.

PROOF. Let $T > 0$. It suffices to show that almost surely, (6.8) holds for all $t \in [0, T]$ and all $A \in \mathcal{B}$. For each $r \in \mathcal{R}$, define the random variable

$$X_T^r = \sup_{C\in\mathcal{C}} \sup_{0\le s\le t\le T} |\bar{\mathcal{L}}^r(s,t)(C) - \lambda^r(t-s)\breve{\vartheta}^r(C)|. \tag{6.9}$$

By Lemma 5.1, $X_T^q \xrightarrow{\mathbf{P}^q} 0$ as $q\to\infty$. Because the limit is deterministic, this convergence is joint with the convergence $\bar{\mathcal{Z}}^q(\cdot) \Rightarrow \mathcal{Z}(\cdot)$. Using the Skorohod representation theorem, assume without loss of generality that $\{\bar{\mathcal{Z}}^q(\cdot), X_T^q\}$ and $\mathcal{Z}(\cdot)$ are defined on a common probability space such that

$$(\bar{\mathcal{Z}}^q(\cdot), X_T^q) \to (\mathcal{Z}(\cdot), 0) \quad \text{almost surely.} \tag{6.10}$$

The conclusions of Lemmas 6.1 and 6.2 hold almost surely as well. Assume for the remainder of the proof that all random objects are evaluated on the event of probability one such that $\mathcal{Z}(\cdot)$ is continuous and such that (6.1), (6.5), and (6.10) hold.

Fix $t\in[0,T]$ and $C\in\mathcal{C}$. An extension to all Borel sets $A\in\mathcal{B}$ will be made at the end. For each q, (5.10) yields

$$\bar{\mathcal{Z}}^q(t)(C) = \bar{\mathcal{Z}}^q(0)(C + (\bar{S}^q(0,t),t)) + \frac{1}{q}\sum_{i=1}^{q\bar{E}^q(t)} 1_C^+(\bar{B}_i^q(t), \bar{D}_i^q(t)). \tag{6.11}$$

We will obtain (6.8) from (6.11) by letting $q\to\infty$. The convergence in the first component of (6.10) is in the Skorohod J_1-topology on $\mathbf{D}([0,\infty),\mathbf{M})$. However, because $\mathcal{Z}(\cdot)$ is continuous,

$$\bar{\mathcal{Z}}^q(s) \xrightarrow{\mathrm{w}} \mathcal{Z}(s), \quad \text{for all } s\in[0,t]. \tag{6.12}$$

Because $\bar{Z}^q(\cdot) = \langle 1, \bar{\mathcal{Z}}^q(\cdot)\rangle$ and $Z(\cdot) = \langle 1, \mathcal{Z}(\cdot)\rangle$, this implies that

$$\lim_{q\to\infty} \left\|\bar{Z}^q(\cdot) - Z(\cdot)\right\|_t = 0. \tag{6.13}$$

For all $t\ge v\ge u>0$, (6.1) implies that $\inf_{s\in[u,v]} Z(s) > 0$, and so the bounded convergence theorem yields

$$\begin{aligned}\lim_{q\to\infty} \bar{S}^q(u,v) &= \lim_{q\to\infty}\int_u^v \frac{1}{\bar{Z}^q(s)}\,\mathrm{d}s\\ &= \int_u^v \frac{1}{Z(s)}\,\mathrm{d}s\\ &= S(u,v).\end{aligned} \tag{6.14}$$

If $Z(0)\ne 0$, then (6.14) holds for $u=0$ as well, because then $\inf_{s\in[0,v]} Z(s) > 0$. If $Z(0)=0$, then $S(0,v)=\infty$ and $\bar{S}^q(0,v)\to\infty$ as $q\to\infty$.

Suppose that $Z(0)\ne 0$ and let $\varepsilon>0$. By (6.14), there exists a $q_\varepsilon\in\mathcal{R}$ such that $\bar{S}^q(0,t)\in((\bar{S}(0,t)-\varepsilon)^+, \bar{S}(0,t)+\varepsilon)$ for $q>q_\varepsilon$. Deduce from the shape of the set C, (6.12), and (6.5) that

$$\limsup_{q\to\infty} \bar{\mathcal{Z}}^q(0)(C+(\bar{S}^q(0,t),t)) \le \bar{\mathcal{Z}}(0)(C+((\bar{S}(0,t)-\varepsilon)^+,t)),$$

$$\liminf_{q\to\infty} \bar{\mathcal{Z}}^q(0)(C+(\bar{S}^q(0,t),t)) \ge \bar{\mathcal{Z}}(0)(C+(\bar{S}(0,t)+\varepsilon,t)).$$

By (6.5), letting $\varepsilon\to 0$ yields

$$\lim_{q\to\infty} \bar{\mathcal{Z}}^q(0)(C+(\bar{S}^q(0,t),t)) = \bar{\mathcal{Z}}(0)(C+(\bar{S}(0,t),t)). \tag{6.15}$$

If $Z(0)=0$, then (6.15) holds trivially because the left side is bounded above by $\lim_{q\to\infty}\langle 1, \bar{\mathcal{Z}}^q(0)\rangle = 0$ by (6.13). Combining with (6.12) and (6.5) for $\bar{\mathcal{Z}}^q(t)$ implies that, as $q\to\infty$,

$$\bar{\mathcal{Z}}^q(t)(C) - \bar{\mathcal{Z}}^q(0)(C+(\bar{S}^q(0,t),t)) \to \mathcal{Z}(t)(C) - \mathcal{Z}(0)(C+(S(0,t),t)).$$

Let I^q denote the second right-hand term in (6.11). Let $\delta>0$ and let $\eta\in(0,t)$. Because $\bar{S}^q(s,t)$ is nonincreasing in s and $S(\cdot,t)$ is continuous on $[\eta,t]$, (6.14) implies that $\bar{S}^q(\cdot,t)\to S(\cdot,t)$ uniformly on $[\eta,t]$. That is, there exists $q_\delta\in\mathcal{R}$ such that

$$\sup_{s\in[\eta,t]} |\bar{S}^q(s,t) - S(s,t)| \le \delta, \quad \text{for all } q>q_\delta. \tag{6.16}$$

Let $D_\vartheta(\mathscr{C}) = \{C \in \mathscr{C}\colon\ \vartheta(\partial_C) \neq 0\}$. Note that $D_\vartheta(\mathscr{C})$ is countable because $\vartheta(\cdot \times \overline{\mathbb{R}}_+)$ and $\vartheta(\overline{\mathbb{R}}_+ \times \cdot)$ are probability measures. Because $Z(u) > 0$ for all $u \in [\eta, t]$, the function $S(s,t)$ is strictly decreasing in s on $[\eta, t]$. Thus,

$$D_\vartheta(S) = \{s \in [\eta, t]\colon\ C + (S(s,t) \pm 2\delta, t - s) \in D_\vartheta(\mathscr{C})\}$$

is also countable. For each integer $N > 1$, let $\eta = t_0^N < t_1^N < \cdots < t_N^N = t$ be a partition of $[\eta, t]$ such that $t_j^N \notin D_\vartheta(S)$ for all $j = 1, \ldots, N-1$, and such that $\max_{j \le N-1}(t_{j+1}^N - t_j^N) \to 0$ as $N \to \infty$. Then,

$$I^q = \frac{1}{q}\sum_{i=1}^{q\bar{E}^q(\eta)} 1_C^+(\bar{B}_i^q(t), \bar{D}_i^q(t)) + \sum_{j=0}^{N-1}\frac{1}{q}\sum_{i=q\bar{E}^q(t_j^N)+1}^{q\bar{E}^q(t_{j+1}^N)} 1_C^+(\bar{B}_i^q(t), \bar{D}_i^q(t)).$$

Note that the first right-hand term is bounded above by $\bar{\mathscr{L}}^q(0,\eta)(\overline{\mathbb{R}}_+^2)$. Suppose that $t_j^N \le U_i^q q^{-1} \le t_{j+1}^N$, for some $q > q_\delta$, some $j \le N-1$, and some $i \in \{q\bar{E}^q(\eta)+1, \ldots, q\bar{E}^q(t)\}$. Then, by (6.16),

$$S(t_{j+1}^N, t) - \delta \le \bar{S}^q(U_i^q q^{-1}, t) \le S(t_j^N, t) + \delta. \tag{6.17}$$

By definition,

$$(\bar{B}_i^q(t), \bar{D}_i^q(t)) = (B_i^q - \bar{S}^q(U_i^q q^{-1}, t), D_i^q q^{-1} - (t - U_i^q q^{-1})).$$

So, for $q > q_\delta$, (6.17) and the inequalities $1_C(\cdot - \delta, \cdot) \le 1_C^+(\cdot,\cdot) \le 1_C(\cdot + \delta, \cdot)$ yield

$$1_C^+(\bar{B}_i^q(t), \bar{D}_i^q(t)) \ge 1_C(B_i^q - (S(t_j^N, t) + 2\delta), D_i^q q^{-1} - (t - t_j^N));$$
$$1_C^+(\bar{B}_i^q(t), \bar{D}_i^q(t)) \le 1_C(B_i^q - (S(t_{j+1}^N, t) - 2\delta), D_i^q q^{-1} - (t - t_{j+1}^N)).$$

This yields, for $q > q_\delta$,

$$I^q \ge \sum_{j=0}^{N-1}\frac{1}{q}\sum_{i=q\bar{E}^q(t_j^N)+1}^{q\bar{E}^q(t_{j+1}^N)} 1_C(B_i^q - (S(t_j^N, t) + 2\delta), D_i^q q^{-1} - (t - t_j^N));$$

$$I^q \le \bar{\mathscr{L}}^q(0,\eta)(\overline{\mathbb{R}}_+^2) + \sum_{j=0}^{N-1}\frac{1}{q}\sum_{i=q\bar{E}^q(t_j^N)+1}^{q\bar{E}^q(t_{j+1}^N)} 1_C(B_i^q - (S(t_{j+1}^N, t) - 2\delta), D_i^q q^{-1} - (t - t_{j+1}^N)).$$

Rewrite as

$$\begin{aligned} I^q &\ge \sum_{j=0}^{N-1} \bar{\mathscr{L}}^q(t_j^N, t_{j+1}^N)(C + (S(t_j^N, t) + 2\delta, t - t_j^N)); \\ I^q &\le \bar{\mathscr{L}}^q(0,\eta)(\overline{\mathbb{R}}_+^2) + \sum_{j=0}^{N-1} \bar{\mathscr{L}}^q(t_j^N, t_{j+1}^N)(C + (S(t_{j+1}^N, t) - 2\delta, t - t_{j+1}^N)). \end{aligned} \tag{6.18}$$

By (6.9) and (6.18), $q > q_\delta$ implies that

$$I^q \ge \sum_{j=0}^{N-1}(\lambda^q(t_{j+1}^N - t_j^N)\breve{\vartheta}^q(C + (S(t_j^N, t) + 2\delta, t - t_j^N)) - X_T^q);$$

$$I^q \le \lambda^q\eta + X_T^q + \sum_{j=0}^{N-1}(\lambda^q(t_{j+1}^N - t_j^N)\breve{\vartheta}^q(C + (S(t_{j+1}^N, t) - 2\delta, t - t_{j+1}^N)) + X_T^q).$$

By (6.10) and because $t_j^N \notin D_\vartheta(S)$ for all $j = 1, \ldots, N-1$,

$$\begin{aligned} \liminf_{q\to\infty} I^q &\ge \lambda\sum_{j=0}^{N-1}(t_{j+1}^N - t_j^N)\vartheta(C + (S(t_j^N, t) + 2\delta, t - t_j^N)); \\ \limsup_{q\to\infty} I^q &\le \lambda\eta + \lambda\sum_{j=0}^{N-1}(t_{j+1}^N - t_j^N)\vartheta(C + (S(t_{j+1}^N, t) - 2\delta, t - t_{j+1}^N)). \end{aligned} \tag{6.19}$$

For $s \in [\eta, t]$ such that $s \notin D_\vartheta(S)$, the bounded convergence theorem implies that

$$\lim_{N\to\infty} \sum_{j=0}^{N-1} 1_{[t_j^N, t_{j+1}^N)}(s)\vartheta(C + (S(t_j^N, t) + 2\delta, t - t_j^N)) = \vartheta(C + (S(s,t) + 2\delta, t - s));$$
$$\lim_{N\to\infty} \sum_{j=0}^{N-1} 1_{[t_j^N, t_{j+1}^N)}(s)\vartheta(C + (S(t_{j+1}^N, t) - 2\delta, t - t_{j+1}^N)) = \vartheta(C + (S(s,t) - 2\delta, t - s)). \quad (6.20)$$

Thus, the convergence in (6.20) holds for almost every $s \in [\eta, t)$. Let $N \to \infty$ in (6.19) and conclude from (6.20) and the bounded convergence theorem that

$$\liminf_{q\to\infty} I^q \geq \lambda \int_\eta^t \vartheta(C + (S(s,t) + 2\delta, t - s))\,\mathrm{d}s;$$
$$\limsup_{q\to\infty} I^q \leq \lambda\eta + \lambda \int_\eta^t \vartheta(C + (S(s,t) - 2\delta, t - s))\,\mathrm{d}s. \quad (6.21)$$

Let $\delta \to 0$ in (6.21). Because $D_\vartheta(\mathscr{C})$ is countable, both integrands in (6.21) converge almost everywhere on $[\eta, t]$ to $\vartheta(C + (S(s,t), t - s))$. Thus,

$$\liminf_{q\to\infty} I^q \geq \lambda \int_\eta^t \vartheta(C + (S(s,t), t - s))\,\mathrm{d}s;$$
$$\limsup_{q\to\infty} I^q \leq \lambda\eta + \lambda \int_\eta^t \vartheta(C + (S(s,t), t - s))\,\mathrm{d}s.$$

Let $\eta \to 0$ to conclude that

$$\lim_{q\to\infty} I^q = \lambda \int_0^t \vartheta(C + (S(s,t), t - s))\,\mathrm{d}s.$$

This proves (6.8) for all $t \in [0, T]$ and $C \in \mathscr{C}$. To extend to all $A \in \mathscr{B}$, apply the $\pi\lambda$-argument appearing in §2.3. □

References

[1] Barrer, D. 1957. Queueing with impatient customers and ordered service. *Oper. Res.* **5** 650–656.
[2] Billingsley, P. 1986. *Probability and Measure*, 2nd ed. John Wiley & Sons, New York.
[3] Bonald, T., L. Massoulié. 2001. Impact of fairness on Internet performance. *Proc. ACM Sigmetrics*, Cambridge, MA, 82–91.
[4] Bonald, T., J. Roberts. 2003. Congestion at flow level and the impact of user behaviour. *Comput. Networks* **42** 521–536.
[5] Boots, N., H. Tijms. 1999. A multi-server queueing system with impatient customers. *Management Sci.* **45** 444–448.
[6] Bramson, M. 2005. Stability of networks for max-min fair routing. *13th INFORMS Appl. Probab. Conf.* Ottawa, Canada.
[7] Coffman, E., A. Puhalskii, M. Reiman, P. Wright. 1994. Processor shared buffers with reneging. *Performance Eval.* **19** 25–46.
[8] de Veciana, G., T. Konstantopoulos, T.-J. Lee. 2001. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Trans. Networking* **9**(1) 2–14.
[9] Doytchinov, B., J. Lehoczky, S. Shreve. 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* **11**(2) 332–378.
[10] Gans, N., G. Koole, A. Mandelbaum. 2002. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.
[11] Gromoll, H. C., P. Robert, B. Zwart, R. Bakker. 2006. The impact of reneging in processor sharing queues. *Proc. ACM Sigmetrics* (June), ACM/IFIP WG 7.3, Saint Malo, France.
[12] Gromoll, H. C., Ł. Kruk. 2007. Heavy traffic limit for a processor sharing queue with soft deadlines. *Ann. Appl. Probab.* **17** 1049–1101.
[13] Gromoll, H. C., R. J. Williams. 2006. Fluid approximation for an Internet congestion control model with fair bandwidth sharing and general document size distributions. Preprint.
[14] Guillemin, F., P. Robert, B. Zwart. 2004. Tail asymptotics for processor-sharing queues. *Adv. Appl. Probab.* **36** 525–543.
[15] Hale, J., S. Verduyn Lunel. 1993. *An Introduction to Functional Differential Equations*. Springer Verlag, New York.
[16] Jean-Marie, A., P. Robert. 1994. On the transient behavior of some single server queues. *Queueing Systems—Theory Appl.* **17** 129–136.
[17] Kallenberg, O. 1986. *Random Measures*. Academic Press, New York.
[18] Kelly, F. P., R. J. Williams. 2004. Fluid model for a network operating under a fair bandwidth sharing policy. *Ann. Appl. Probab.* **14** 1055–1083.
[19] Key, P., L. Massoulié, A. Bain, F. Kelly. 2004. Fair internet traffic integration: Network flow models and analysis. *Ann. Telecomm.* **59** 1338–1352.
[20] Kruk, Ł., J. Lehoczky, S. Shreve. 2003. Second order approximation for the customer time in queue distribution under the FIFO service discipline. *Ann. Univ. Marie Curie-Skłodowska, Sect. AI Inform* **1** 37–48.
[21] Kruk, Ł., J. Lehoczky, S. Shreve. 2006. Accuracy of state space collapse for earliest-deadline-first queues. *Ann. Appl. Probab.* **13**(2) 516–581.

[22] Massoulié, L., J. Roberts. 1999. Bandwidth sharing: Objectives and algorithms. *INFOCOM'99. 18th Annual Joint Conf. IEEE Comput. Comm. Societies*, New York, 1395–1403.
[23] Mo, J., J. Walrand. 2000. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Networking* **8**(5) 556–567.
[24] Puha, A. L., A. L. Stolyar, R. J. Williams. 2006. The fluid limit of an overloaded processor sharing queue. *Math. Oper. Res.* **31**(2) 316–350.
[25] Roberts, J., L. Massoulié. 2000. Bandwidth sharing and admission control for elastic traffic. *Telecomm. Systems* **15** 185–201.
[26] Stanford, R. E. 1979. Reneging phenomena in single channel queues. *Math. Oper. Res.* **4** 162–178.
[27] Stanford, R. E. 1990. On queues with impatience. *Adv. Appl. Probab.* **22**(3) 768–769.
[28] van der Vaart, A., J. A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.
[29] Ward, A., P. Glynn. 2003. A diffusion approximation for a markovian queue with reneging. *Queueing Systems* **43** 103–128.