Fluid Limits for Processor Sharing Queues with Impatience

Christian Gromoll, Philippe Robert, Bert Zwart

FLUID LIMITS FOR PROCESSOR SHARING QUEUES WITH IMPATIENCE

H. CHRISTIAN GROMOLL*, PHILIPPE ROBERT, AND BERT ZWART†

ABSTRACT. 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 model captures many essential features of the underlying stochastic model, and it is used to analyze the impact of impatience in processor sharing queues.

CONTENTS

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 [29] and Kelly and Williams [18] and the 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.

1991 *Mathematics Subject Classification.* Primary 60K25, 60K30, 60G57, 60F17. Secondary 90B15, 90B22.

Key words and phrases. Processor Sharing. Queues with Impatience. Measure Valued Process. Fluid Limits. Delay-Differential Equations. Empirical Processes.

*Research supported in part by an NSF Mathematical Sciences Postdoctoral Research Fellowship, a European Union Marie Curie Postdoctoral Research Fellowship, and EURANDOM.

†Research supported by an NWO-VENI grant.

For example, the timeout of a 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 larger 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. Since this partial service is wasted, impatience may create a 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 [31] is a survey of the literature in this domain (see also Stanford [30] and Boots and Tijms [5]). This body of work focuses primarily on exact performance analysis. Ward and Glynn [33] have recently obtained a diffusion approximation for single channel queues. There are also various studies of multi-server queues with abandonments, motivated by call center applications; see the survey by 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. 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], Kruk *et al.* [20, 21], and Yeung and Lehoczky [34] investigate the heavy traffic behavior of various systems using the Earliest Deadline First and FIFO policies. Gromoll and Kruk [12] describes the heavy traffic behavior of a PS queue incorporating a fairly general structure of soft deadlines.

For PS queues with impatience however, 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 by 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 by 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 an analogue representation 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, it is shown that, with a convenient scaling, a family of measure valued processes associated with $(\mathcal{Z}(t))$ is tight and converges in distribution to some $(\zeta(t))$. For $t \geq 0$, $\zeta(t)$ is a nonnegative measure on $(0,\infty] \times (0,\infty]$, it is the limit in distribution of the sequence of random points describing the queue.

This fluid limit is characterized as the solution of a functional Equation (2.8) which can be viewed as a time changed functional differential equation.

The overloaded case $\rho > 1$, which forms our main focus, presents a nontrivial and quite 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 (3.2). Moreover, the fluid steady state, i.e. the limit of $\zeta(t)$ as t goes to infinity, is a distribution on $(0,\infty] \times (0,\infty]$ which has a simple expression (2.11) in terms of z_∞.

These results give also a significant insight on the qualitative properties of PS queues with impatience. An interpretation of the fixed point equation (3.2) is given and used to analyze the total number of jobs in the system and to estimate the fraction of jobs that renege. The impact of the variability of the service times and of the lead times and other properties of this queue are extensively investigated in Gromoll *et al.* [11].

In contrast to the models studied previously in this domain, the service discipline considered here is *not work conserving*. For this reason, analysis of the fluid model is more intricate. This is an important difference from earlier work on standard PS queues where the fact that the workload process coincides with that of FIFO discipline was *a crucial ingredient* in the proof of the key results. A different approach to prove existence, uniqueness, and convergence to steady state of fluid model solutions is proposed. It is shown that there exists a *maximal* fluid model solution and by using monotonicity arguments, the properties of the fluid limits can be investigated under quite general assumptions.

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

2. Model Description and Results

This section gives a detailed description of the stochastic processes associated to this queue and a summary of the 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 $E(\cdot)$, $\{B_i, D_i\}_{i=1}^{\infty}$ describing respectively the process of arrivals and the services and the deadlines of the customers, and a random initial condition specifying the state of the system at time 0. All random objects are defined on a probability space $(\Omega, \mathscr{F}, \mathbf{P})$ with expectation operator $\mathbf{E}(\cdot)$.

The *exogenous arrival process* $(E(t), t \geq 0)$ has rate λ>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; 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. Since it 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. Therefore, the random variable (B_i, D_i) has values in the space $\overline{\mathbb{R}}_+^2 = [0,\infty] \times [0,\infty]$. Here, $\overline{\mathbb{R}}_+ = [0,\infty]$ is the usual compactification of $\mathcal{R}_+$ with the arithmetic extensions $x + \infty = \infty$ for all $x \in \overline{\mathbb{R}}_+$, $x \cdot \infty = \infty$ for $x > 0$ and $0 \cdot \infty = 0$. 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 services and deadlines $\{B_i, D_i\}_{i=1}^{\infty}$ 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

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

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). Let $\rho = \lambda \mathbf{E}[B]$ denote the *traffic intensity* of the system. It is assumed throughout that $\rho > 1$, that is, the server is nominally overloaded. In this way, the classical PS queue, the infinite server queue, and mixtures of the two are special cases of this model. (For example the $GI/GI/\infty$ queue corresponds to the case when service times are equal to infinity.)

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 non-negative, 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 i.i.d. random variables taking values in $\{(0,\infty] \times (0,\infty]\} \setminus (\infty,\infty)$ almost surely. A generic random element of $\overline{\mathbb{R}}_+^2$ distributed as (B_0^0, D_0^0) will be denoted by (B^0, D^0). 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 $S(t)$ denotes the *cumulative service time per job* provided by the server up to time t. Because of the processor sharing policy, the quantity $S(t)$ is given by

$$S(t) = \int_0^t \frac{1}{Z(s)}\, \mathrm{d}s, \tag{2.1}$$

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 $t \geq s$, at time t it has received the cumulative amount of processing time $S(t) - S(s)$.

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

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

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

$$D_i(t) = (U_i + D_i - t)^+, \qquad \text{and} \qquad D_j^0(t) = (D_j^0 - t)^+. \tag{2.2}$$

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.

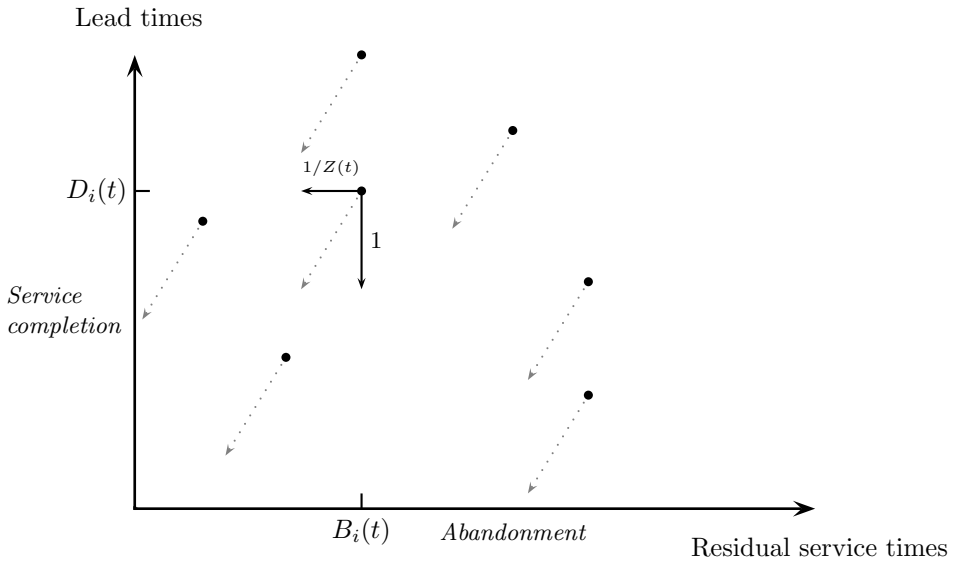

FIGURE 1. Dynamics of the measure valued process $\mathcal{Z}(\cdot)$

Job i will depart the system either when its service requirement is fulfilled or when its deadline arrives, it will leave the system at time

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

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

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

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

$$Z(t) = \langle 1, \mathcal{Z}(t) \rangle,$$

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 $f : \overline{\mathbb{R}}_+^2 \to \mathbb{R}$.

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 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 $\mathbf{M_1}$ denote the space of finite non-negative Borel measures on $\overline{\mathbb{R}}_+^2$, endowed with the topology of weak convergence: $\zeta_n \xrightarrow{\mathbf{w}} \zeta$ in $\mathbf{M_1}$ if and only if $\langle f, \zeta_n \rangle \to \langle f, \zeta \rangle$ for all continuous functions $f : \overline{\mathbb{R}}_+^2 \to \mathbb{R}$ (recall that $\overline{\mathbb{R}}_+^2$ is compact for the induced topology). Let $\mathbf{D}([0,\infty), \mathbf{M_1})$ denote the space of *càdlàg* paths in $\mathbf{M_1}$, endowed with the Skorohod J_1-topology. Then, for $t \geq 0$, $\mathcal{Z}(t)$ is a random element of $\mathbf{M_1}$ for each $t \geq 0$, and $\mathcal{Z}(\cdot)$ is a random element of $\mathbf{D}([0,\infty), \mathbf{M_1})$.

It is clear that, given stochastic primitives $(E(\cdot), \{B_i, D_i\}_{i=1}^{\infty})$ and the initial condition $\mathcal{Z}(0)$, the equation (2.1) uniquely determines 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 $A \in \mathscr{B}$, and all $t \geq 0$,

$$\mathcal{Z}(t)(A) = \mathcal{Z}(0)\left(A + (S(t), t)\right) + \sum_{i=1}^{E(t)} 1_A^+\left(B_i(t), D_i(t)\right), \tag{2.3}$$

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

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

This equation plays a crucial rôle in determining fluid limits for the model.

2.2. **A Fluid Scaling.** A sequence of renormalized stochastic processes $(\overline{\mathcal{Z}}^r(t))$ associated to the solution of the evolution equation (2.3) is introduced. The limits of $(\overline{\mathcal{Z}}^r(t))$ will give the fluid limits of this queue.

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 Section 2.1. That is, for each $r \in \mathcal{R}$, there are stochastic primitives $(E^r(\cdot), \{B_i^r, D_i^r\}_{i=1}^{\infty})$ with associated data λ^r and ϑ^r, and an initial condition $\mathcal{Z}^r(0)$ which give 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 $\{B_i^{0r}(\cdot), D_i^{0r}(\cdot)\}$. Each model is defined on a probability space $(\Omega^r, \mathscr{F}^r, \mathbf{P}^r)$ with expectation operator $\mathbf{E}^r(\cdot)$.

A fluid scaling is applied to each model in the sequence. To obtain non-trivial scaling limits, initial lead times $\{D_i^r\}$ will be assumed to be of order r. For each $r \in \mathcal{R}$, let $\breve{\vartheta}^r \in \mathbf{M_1}$ be the probability measure defined by

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

for all Borelian subsets $F, G \in \mathscr{B}$ with the notation $rG = \{r \cdot g : g \in G\}$. Note that if $\{B_i^r, D_i^r\} = \{B_i, rD_i\}$ for some sequence $\{B_i, D_i\}$, then $\breve{\vartheta}^r$ 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 $\overline{\mathcal{Z}}^r(t) \in \mathbf{M}$ such that

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

for all Borelian subsets $F, G \in \mathscr{B}$. 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

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

The following asymptotic assumptions are needed. Let $(\lambda, \vartheta, \zeta_0)$ be fluid model data satisfying the assumptions of Section 2.3. Assume that as $r \to \infty$,

$$\overline{E}^r(\cdot) \to \lambda(\cdot), \tag{2.4}$$

$$\breve{\vartheta}^r \xrightarrow{\mathbf{w}} \vartheta, \tag{2.5}$$

$$\overline{\mathcal{Z}}^r(0) \xrightarrow{\mathbf{w}} \zeta_0, \quad \text{in distribution,} \tag{2.6}$$

in particular, Assumption (2.4) implies that $\lambda^r \to \lambda$ holds.

2.3. **Fluid model.** A deterministic fluid model satisfying dynamic equations analogous to (2.3) is introduced. It will be shown later that these equations can be obtained as limits of Equation (2.3) under an appropriate scaling procedure.

Let $\rho > 1$, and $\zeta_0 \in \mathbf{M_1}$ be a measure on $\overline{\mathbb{R}}_+^2$ such that the projections $\zeta_0(\cdot \times \overline{\mathbb{R}}_+)$ and $\zeta_0(\overline{\mathbb{R}}_+ \times \cdot)$ are free of atoms in $[0, \infty)$ and $z_0 = \zeta_0(0,0)$ is the total mass of ζ_0.

Definition 2.1. *A* measure valued fluid model solution *for the data* $(\lambda, \vartheta, \zeta_0)$ *is a continuous function* $\zeta(\cdot) : [0, \infty) \to \mathbf{M_1}$ *such that*

(i) $\inf_{t>a} z(t) > 0$ *for all* $a > 0$,
(ii) *for all* $C \in \mathcal{C}$ *and* $t \geq 0$,

$$\zeta(t)(C) = \zeta_0(C + (S(0,t), t)) + \lambda \int_0^t \vartheta(C + (S(s,t), t-s))\,\mathrm{d}s, \tag{2.7}$$

where $S(u,v) = \int_u^v 1/z(s)\,\mathrm{d}s$ *for all* $v \geq u \geq 0$ *and* $z(\cdot)$ *is the total mass function is* $z(\cdot) = \langle 1, \zeta(\cdot)\rangle$. *The function* $z(\cdot)$ *is simply called a* fluid model solution *for* $(\lambda, \vartheta, \zeta_0)$.

Note that $S(0,t)$ may be equal to $+\infty$ if $\zeta_0 \equiv 0$, i.e. $z(0) = 0$. Both right hand side terms in (2.7) are still well defined in this case, and the first term equals zero.

The class of corner sets is defined as

$$\mathcal{C} = \{[x, \infty) \times [y, \infty) : x, y \in \mathbb{R}_+\} \cup \left\{[x, \infty] \times [y, \infty] : x, y \in \overline{\mathbb{R}}_+\right\}.$$

The sets from the class $\mathcal{C}$ will be used to describe the evolution of fluid model solutions. Since each $C = [x, \infty] \times [y, \infty] \in \mathcal{C}$ is characterized by the coordinates (x, y) of its corner, it is convenient to use the notation $\mu(x,y) \stackrel{\text{def.}}{=} \mu([x,\infty] \times [y,\infty])$ for any $\mu \in \mathbf{M_1}$. If $z_0 > 0$, for this class of subsets Equation (2.7) can then be rewritten as follows: for each $x, y \geq 0$ and $t \geq 0$,

$$\zeta(t)(x,y) = z_0 \mathbf{P}\left(B^0 > x + S(0,t), D^0 > y + t\right) + \lambda \int_0^t \mathbf{P}\left(B > x + S(s,t), D > y + t - s\right) \mathrm{d}s, \tag{2.8}$$

Since $z(t) = \zeta(t)(0,0)$, the fluid model solution $z(\cdot)$ satisfies the following equation: for each $t \geq 0$,

$$z(t) = z_0 \mathbf{P}\left(B^0 > S(0,t); D^0 > t\right) + \lambda \int_0^t \mathbf{P}\left(B > S(s,t); D > t-s\right) \mathrm{d}s. \tag{2.9}$$

It will be proved that the fluid model defined above is the limit in distribution of the rescaled processes $\{\overline{\mathcal{Z}}^r(\cdot) : r \in \mathcal{R}\}$ introduced in Section 2.1. The measure valued fluid model solution $\zeta(\cdot)$ corresponds to the measure valued state descriptor $\mathcal{Z}(\cdot)$, and the fluid model solution $z(\cdot)$ is the limit of the queue length process $Z(\cdot)$. The main result concerning the convergence of $\{\overline{\mathcal{Z}}^r(\cdot) : r \in \mathcal{R}\}$ is the following theorem.

Theorem 2.2. *The sequence $\{\overline{\mathcal{Z}}^r(\cdot) : r \in \mathcal{R}\}$ 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* (2.10) *holds, then $\overline{\mathcal{Z}}^r(\cdot)$ converges in distribution, as $r \to \infty$, to the unique measure valued fluid model solution $\zeta(\cdot)$.*

This theorem is proved in Section 5 and 6.

2.4. **Some 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 of the queue can be obtained. Let $(\lambda, \vartheta, \zeta_0)$ satisfy the assumptions of Section 2.3.

The first result establishes the uniqueness of fluid model solutions under a Lipschitz condition on the initial condition ζ_0.

Theorem 2.3. *Suppose there exists a finite constant L such that*

$$\zeta_0(A \times [y, y']) \leq L|y' - y| \tag{2.10}$$

for all Borel sets A of $\overline{\mathbb{R}}$ and all $y' > y \geq 0$. Then (2.8) *and* (2.9) *have a unique solution.*

The second theorem analyzes the equilibrium of the fluid model, i.e. the behavior at infinity of the solution of Equation (2.9).

Theorem 2.4. *Suppose that $\lambda \mathbf{E}\left[B 1_{\{D=\infty\}}\right] < 1$ and $\mathbf{E}\left[\min\{B, D\}\right] < \infty$. Then any solution $t \to z(t)$ of* (2.9) *converges at infinity to the unique positive solution z_∞ of the fixed point equation*

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

Moreover, any solution $(\zeta(t))$ of Equation (2.8) *converges to the measure $\zeta_\infty \in \mathbf{M_1}$, defined by*

$$\zeta_\infty(x,y) = \lambda \int_0^\infty \mathbf{P}\left(B > x + \frac{t}{z_\infty}, D > y + t\right) \mathrm{d}t \\ = \mathbf{E}\left[\min\{z_\infty (B-x)^+, (D-y)^+\}\right], \tag{2.11}$$

for $x, y \geq 0$.

These theorems are proved in Section 3. The simple fixed point equation stated in this theorem is used to analyze the qualitative behavior of the queue, see Section 4. Note that the expression of the distribution of points ζ_∞ describing the asymptotic behavior of this queue has a simple expression in terms of the solution z_∞ of the fixed point equation.

3. Some 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, i.e. such that

$$z(t) = z_0\mathbf{P}\left(B^0 > S(0,t); D^0 > t\right) + \lambda \int_0^t \mathbf{P}\left(B > S(s,t); D > t-s\right)\,\mathrm{d}s. \tag{2.9}$$

If $z_0 > 0$, define

$$\widetilde{S}(t) = \inf\{s : S(0,s) \geq t\},$$

since $z(t) \leq \lambda t + z_0$, $S(0,t) \to \infty$ as $t \to \infty$, implying that $\widetilde{S}(t)$ is well defined for all t. In addition, $\widetilde{S}(t) < \infty$ for all t if $z_0 > 0$, which follows from property (i) and continuity of the fluid model solution $z(\cdot)$.

Define $\widetilde{z}(t) = z(\widetilde{S}(t))$, then $(\widetilde{z}(t))$ satisfies the equation

$$\begin{aligned}\widetilde{z}(t) = z_0\mathbf{P}\left(B^0 \geq t; D^0 \geq \widetilde{S}(t)\right) \\ + \lambda \int_0^t \widetilde{z}(u)\mathbf{P}\left(B \geq t-u; D \geq \widetilde{S}(t) - \widetilde{S}(u)\right)\,\mathrm{d}u.\end{aligned} \tag{3.1}$$

We next introduce the concept of a *shifted* fluid model solution: For $t_0 > 0$ define $z(t_0,t) = z(t_0+t)$, and define $S(t_0,u,v) = \int_u^v 1/z(t_0,r)\mathrm{d}r$.

Property 3.1. *If $z(\cdot)$ is a solution of (2.9), then $z(t_0,\cdot) = z(t_0+\cdot)$ satisfies*

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

Proof. Note first that by definition, and since $S(t_0,0,t) = S(t_0,t_0+t)$,

$$\begin{aligned}\zeta(t_0)(S(t_0,0,t),t) &= z_0\mathbf{P}\left(B^0 \geq S(t_0,0,t) + S(0,t_0), D^0 \geq t_0+t\right) \\ &\quad + \lambda\int_0^{t_0} \mathbf{P}\left(B \geq S(t_0-s,t_0; D \geq s+t) + S(t_0,0,t)\right)\mathrm{d}s \\ &= z_0\mathbf{P}\left(B^0 \geq S(0,t_0+t); D^0 \geq t_0+t\right) \\ &\quad + \lambda\int_0^{t_0} \mathbf{P}\left(B \geq S(t_0-s,t_0+t); D \geq s+t\right)\mathrm{d}s. \\ &= z_0\mathbf{P}\left(B^0 \geq S(0,t_0+t); D^0 \geq t_0+t\right) \\ &\quad + \lambda\int_0^{t_0} \mathbf{P}\left(B \geq S(s,t_0+t); D \geq t_0+t-s\right)\mathrm{d}s.\end{aligned}$$

We apply this expression as follows:

$$\begin{aligned}z(t_0,t) &= z(t_0+t) \\ &= z_0\mathbf{P}\left(B^0 \geq S(0,t_0+t); D^0 \geq t_0+t\right) \\ &\quad + \lambda\int_0^{t_0+t} \mathbf{P}\left(B \geq S(s,t_0+t); D \geq t_0+t-s\right)\mathrm{d}s \\ &= \zeta(t_0)(S(t_0,0,t),t) + \lambda\int_{t_0}^{t_0+t} \mathbf{P}\left(B \geq S(s,t_0+t); D \geq t_0+t-s\right)\mathrm{d}s.\end{aligned}$$

The change of variables $y = s - t_0$ and the identity $S(t_0 + y, t_0 + t) = S(t_0, y, t)$ give the result. □

The next proposition shows that continuity of fluid model solutions is a consequence of properties (i) and (ii).

Lemma 3.2. *Let the distribution of (B^0, D^0) be free of atoms. Then any solution $z(t), t \geq 0$ to (2.9) satisfying $\inf_{t>a} z(t) > 0$ for all $a > 0$ is continuous.*

Proof. The function $t \to S(0,t)$ is continuous and so is $t \to \mathbf{P}\left(B^0 > S(0,t); D^0 > t\right)$ since the distribution of (B^0, D^0) has no atom. The first term of the right hand side of Equation (2.9) is a continuous function of t. Concerning the second term of Equation (2.9), by monotonicity the integrand $t \to \mathbf{P}\left(B > S(s,t); D > t - s\right)$ is continuous almost everywhere on $\mathbb{R}_+$, hence its integral is a continuous function of t by Lebesgue's Theorem. The lemma is proved. □

3.1. **A Maximal Solution.** An important monotonicity property of fluid model solutions is proved in this section.

Proposition 3.3. *If $\lambda \mathbf{E}\left[B1_{\{D=\infty\}}\right] < 1$ and $\mathbf{E}\left[\min\{B, D\}\right] < \infty$, then any fluid model solution is bounded.*

Proof. Note first that, since $\mathbf{E}\left[\min\{B, D\}\right] < \infty$, also $\mathbf{E}\left[\min\{B, aD\}\right] < \infty$ 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 $u \in [0, t]$. Since $S(u, s) \geq (u - s)/\|z\|_t$, it holds that

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

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

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

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

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

which contradicts the assumption $\lambda \mathbf{E}\left[B1_{\{D=\infty\}}\right] < 1$. We conclude that $\|z\|_t$ converges to some finite constant M which implies the assertion. □

Proposition 3.4 (Maximal Fluid Solution)**.** *For $z_0 > 0$, there exists a fluid model solution $z^*(\cdot)$ starting form z_0 which is maximal, i.e. for any fluid model solution $z(\cdot)$ such that $z(0) = z_0$, 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 $z^n(\cdot), n \geq 0$, by $z^0(t) = z_0 + \lambda t$, $S^n(u, v) = \int_u^v (1/z^n(r))\mathrm{d}r$ and

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

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 now that $z^n(t) \leq z^{n-1}(t)$. Then $S^n(u, v) \geq S^{n-1}(u, v)$, and, using the

fact that tail probabilities are non-increasing,

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

which equals $z^n(t)$.

Since $z^n(t)$ is decreasing in n and non-negative for all n there exists a function $z^*(t)$ such that $z^*(t) = \lim_{n\to\infty} z^n(t)$. By the definition of $z^n(t)$, we see that $z^*(t)$ satisfies (2.9).

Furthermore, we have $z(t) \leq z^*(t)$ for any given fluid model solution $z(\cdot)$. This is true because $z(t) \leq z^0(t)$, and using an inductive argument as above, $z(t) \leq z^n(t)$ for every n. Since we know that at least one fluid model solution exists, it follows that $\inf_{t>a} z^*(t) > 0$ for every $a > 0$. By Lemma 3.2, it follows that $z^*(t) > 0$ is continuous. We conclude that $z^*(\cdot)$ is indeed a fluid model solution. □

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

Proposition 3.5. *If* $\lambda\mathbf{E}\left[B1_{\{D=\infty\}}\right] <1$, $\mathbf{E}\left[\min\{B,D\}\right] <\infty$ *and* $\rho>1$, *the equation*

$$z_\infty = \lambda\mathbf{E}\left[\min\{z_\infty B, D\}\right], \tag{3.2}$$

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

Proof. The function $f : a \to \lambda\mathbf{E}\left[\min\{B,aD\}\right]$ is non-decreasing and concave on $[0,+\infty)$, note that $f(a) = \lambda\mathbf{E}\left[\min\{B,aD\}1_{\{D<+\infty\}}\right] + \lambda\mathbf{E}\left[B1_{\{D=+\infty\}}\right]$ for $a > 0$, therefore $f(0+) = \lambda\mathbf{E}\left[B1_{\{D=+\infty\}}\right] <1$ and $f(a)$ converges to $\lambda\mathbf{E}\left[B\right] >1$ as a goes to infinity. By continuity of f, there exists a_0, $0<a_0<+\infty$ such that $f(a) = 1$. The concavity and the monotonicity imply that such a a_0 is unique, otherwise f should be constant equal to 1 after a_0, but this is impossible since f converges to $\lambda\mathbf{E}\left[B\right] >1$ at infinity. The quantity $1/a_0$ is then the unique solution of Equation (3.2). □

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 3.6. *If* $z(\cdot)$ *is a fluid model solution, under the conditions*

$$\lambda\mathbf{E}\left[B1_{\{D=\infty\}}\right] < 1, \mathbf{E}\left[\min\{B,D\}\right] < \infty \text{ and } \rho > 1,$$

then $z(t)\to z_\infty$ *as* $t\to+\infty$, *where* z_∞ *the unique positive solution of the fixed point equation (3.2).*

Proof. It suffices to show $\bar{z} = \limsup_{t\to\infty} z^*(t) \leq z_\infty$ and $\underline{z} = \liminf_{t\to\infty} z^*(t) \geq z_\infty$. We start with the former. We know that $\bar{z} < \infty$ from Proposition 3.3. For any $\varepsilon > 0$ there exists a t_ε such that $z(t) \leq \bar{z} + \varepsilon$. We see that, for $t > t_\varepsilon$,

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

Taking the lim sup on both sides, and noting that $S(s,t) \to \infty$ for any $s \geq 0$, we obtain that

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

The result is valid for every $\varepsilon > 0$. By letting $\varepsilon \downarrow 0$, we obtain that $\bar{z} \leq z_\infty$. The lower bound follows by a similar argument after first noting that $\underline{z} > 0$ since $z(\cdot)$ is a fluid model solution. □

3.3. **Uniqueness of fluid model solutions under non-zero conditions.** The uniqueness of fluid model solutions is, in general, difficult to determine. If one looks at the time-changed version (3.1), and take $\lambda = 0$, one gets an ODE. Uniqueness of solutions to such ODE's can usually only be established by reducing it to some special case or to assume some kind of Lipschitz condition. If $D = \infty$, then (3.1) reduces to a renewal equation for which uniqueness is known to hold. Unfortunately, this reduction is not possible in general, which lead us to use a Lipschitz condition on the distribution function of (B^0, D^0). It is not necessary to assume regularity conditions on the distribution of (B, D). We shall give a direct proof of uniqueness; for related results in the functional analysis literature, we refer to Chapter 2 of Hale & Verduyn Lunel [15].

Theorem 3.7. *Suppose $z_0 > 0$ and $F_0(x, y) = \mathbf{P}\left(B^0 \geq x; D^0 \geq y\right)$ is Lipschitz continuous in Y, i.e. there is a constant L such that for any x, y, y',*

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

Then (2.9) has a unique solution.

Defining $\widetilde{\zeta}(t)(u, v) = \zeta(\widetilde{S}(t))(u, v)$, it can easily be shown that

$$\widetilde{\zeta}(t)(u, v) = z_0 \mathbf{P}\left(B^0 \geq u + t; D^0 \geq v + \widetilde{S}(t)\right) + \lambda \int_0^t \widetilde{z}(s) \mathbf{P}\left(B \geq u + (t - s); D \geq v + \widetilde{S}(t) - \widetilde{S}(s)\right) \mathrm{d}s. \tag{3.3}$$

It is clear that for any u, v, $\widetilde{\zeta}(t)(u, v), t \geq 0$, is completely determined by $\widetilde{z}(t), t \geq 0$ and the initial measure. Thus, uniqueness of $z(t)$ on an interval A carries over to uniqueness of $\zeta(t)(u, v)$ on A.

The idea of the proof is simple: we take a suitable constant $a > 0$ and prove first that uniqueness holds for $\widetilde{z}(t)$ for $[0, a]$. As discussed above, uniqueness carries over to $\widetilde{\zeta}(t)(u, v)$ for $t \in [0, a]$. Using this and the shifted fluid model equation given by Property 5.1, we prove uniqueness for $\widetilde{z}(t)$ on the interval $[a, 2a]$, and so forth. This iterative procedure works if the measure $\widetilde{\zeta}(t)(u, v)$ is Lipschitz for any $0 < t < T$. This is the content of the following lemma.

Lemma 3.8. *For any x, y, y' and for any t we have*

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

Proof. We may take y, y' such that $y \leq y'$. From (3.3) we obtain

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

Noting that $\widetilde{z}(s)\mathrm{d}s = \mathrm{d}\widetilde{S}(s)$ we can rewrite this into

$$z_0 L|y' - y| + \lambda \int_0^{\widetilde{S}(t)} \mathbf{P}\left(r + y \leq D \leq r + y'\right) \mathrm{d}s \tag{3.4}$$

Noting that, for any $\delta > 0$

$$\lambda \int_0^\infty \mathbf{P}\left(y \le D < y + \delta\right) \mathrm{d}y \le \delta,$$

we see that (3.4) can be upper bounded by $(z_0 L + \lambda)|y - y'|$. □

Proof of Theorem 3.7. By the one-to-one correspondence between solutions of (2.9) and (3.1), it suffices to show that (3.1) has a unique solution. Define $a = 1/(2(z_0 L + 4\lambda))$. We first show that (3.1) has a unique solution on the interval $[0, a]$. For that, suppose that there exist two different solutions $\widetilde{z}(t), 0 \le t \le a$ and $h(t), 0 \le t \le a$. Set $\varepsilon = \sup_{0 \le t \le a} |\widetilde{z}(t) - h(t)|$.

Note that for any $0 \le s < t \le a$,

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

Using (3.1) for both z and h, together with the Lipschitz assumption, we obtain, after some simple estimates,

$$|\widetilde{z}(t) - h(t)| \le z_0 \left|\mathbf{P}\left(B^0 \ge t; D^0 \ge \widetilde{S}(t)\right) - \mathbf{P}\left(B^0 \ge t; D^0 \ge H(t)\right)\right|$$
$$+\lambda \int_0^t |\widetilde{z}(s)\mathbf{P}\left(B \ge t-s; D \ge \widetilde{S}(t)-\widetilde{S}(s)\right) - h(s)\mathbf{P}\left(B \ge t-s; D \ge H(t)-H(s)\right)| \mathrm{d}s.$$

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

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

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

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

to obtain (after a change of variable $r = \widetilde{S}(s)$)

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

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

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

which implies that $\varepsilon = 0$, i.e. that $\widetilde{z}(t)$ and $h(t)$ coincide on $[0, a]$. Hence Equation (3.1) has a unique solution in the interval $[0, a]$.

Suppose now that (3.1) has a unique solution on $[0, ka]$ for some $k \ge 1$, and consider the equation

$$\widetilde{z}(ka, t) = \zeta(ka)(t, \widetilde{S}(ka, t)) + \lambda \int_0^t \widetilde{z}(ka, s)\mathbf{P}\left(B \ge t - s; D \ge \widetilde{S}(ka, t) - \widetilde{S}(ka, s)\right) \mathrm{d}s. \tag{3.5}$$

We now show that this equation has a unique solution on $[0, a]$, implying that there exist a unique solution of (3.1) on the interval $[0, (k+1)a]$. Suppose $\widetilde{z}(ka, t)$ and $h(t)$ both satisfy (3.5), and set $\varepsilon = \sup_{t\in[0,a]} |z(ka, t) - h(t)|$. As before, we have

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

Using this, we get as before (using now the Lemma for the first term) that

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

which implies that $\varepsilon = 0$, and that uniqueness of solutions of (3.1) holds on the interval $[0, (k+1)a]$. Iterating this argument completes uniqueness for all t. □

3.4. **Uniqueness starting from zero.** The result in this subsection can be seen as an extension of a result of [28], who considered the case PS queue without impatience, i.e. with $D \equiv +\infty$.

Theorem 3.9. *Let $\varepsilon > 0$. Suppose that (B, D) and a non-increasing function $F_\varepsilon(x, y)$, with $0 \leq F_\varepsilon(x, y) \leq \lambda\varepsilon$ are such that*

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

has a unique solution $z_\varepsilon(t)$ satisfying $\inf_{t>a} z_\varepsilon(t) > 0$ for $a > 0$. Then $z_\varepsilon(t) \to z_0^*(t)$ as $\varepsilon \downarrow 0$.

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

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

From this construction it can be easily shown that $z_\varepsilon^n(t)$ is decreasing in n, and that $z_\varepsilon^n(t) \geq z_0^n(t)$. Since also $z_\varepsilon^*(t) \leq z_\varepsilon^n(t)$, we see that

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

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

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

To prove the other bound, we observe by induction and the properties of F_ε that $z_0^n(t) \leq z_\varepsilon^n(t)$ for every $n \geq 0$. Consequently,

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

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

Uniqueness of fluid model solutions starting from 0 is now a simple corollary.

Corollary 3.10. *Suppose that $z_0 = 0$. Then (3.1) has a unique solution.*

Proof. Let $z(\cdot)$ be a fluid model solution. Define $z_\varepsilon(t) = z(t + \varepsilon)$. Given $z(s), 0 \leq s \leq \varepsilon$, $z_\varepsilon(\cdot)$ satisfies the equation

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

Here (with obvious notation)

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

We see that F_ε is globally Lipschitz in the second coordinate (with Lipschitz constant 1). Consequently, the above equation has a unique fluid model solution in terms of F_ε so that $z_\varepsilon(\cdot)$ is uniquely determined by $z(t), 0 \leq t \leq \varepsilon$. Since $F_\varepsilon(x,y) \leq \lambda\varepsilon$, we see from the previous theorem that $z_\varepsilon(t) \to z_0^*(t)$. But also $z_\varepsilon(t) = z(t+\varepsilon) \to z(t)$, since $z(t)$ is continuous. We conclude that $z(t) = z_0^*(t)$, which implies uniqueness. □

3.5. **Analysis of the measure-valued fluid model.** For any Borel set F of $\overline{\mathbb{R}}_+^2$, the measure valued function $\zeta(\cdot)$ satisfies the equation

$$\zeta(t)(F) = \zeta_0(F + (S(0,t),t)) + \lambda \int_0^t \vartheta(F + (S(s,t), t-s))\mathrm{d}s. \tag{3.6}$$

The properties, which are analogues of properties of $z(\cdot)$, are gathered in the following theorem:

Theorem 3.11. *Let $\zeta(\cdot)$ be a solution of (3.6).*

(i) *Suppose $\rho > 1, \mathbf{E}[\min\{B,D\}] < \infty$ and $\lambda\mathbf{E}[B1_{\{D=\infty\}}] < 1$. As $t \to \infty$, $\zeta(t)$ converges to the limiting measure ζ_∞ defined by*

$$\zeta_\infty([x,\infty] \times [y,\infty]) = \lambda \int_0^\infty \mathbf{P}\left(B - x \geq s/z_\infty, D - y \geq s\right)\mathrm{d}s,$$

where z_∞ is the unique solution of the fixed point equation (3.2).

(ii) *If Condition (2.10) of Theorem 2.3 holds, then Equation* (3.6) *has a unique solution.*

Proof. We know that $z(t) \to z_\infty$ as $t \to \infty$. Consequently, $S(t-t_0,t) \to t_0/z_\infty$ for every $t_0 > 0$. Write for any Borel set F,

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

Number the three terms on the right hand side as I, II, III. By shifting time if necessary, we may assume that $z(0) > 0$. The first term converges to 0. Since $z(0) > 0$, there exists an $\eta > 0$ such that $z(t) \geq \eta$ for all $t \geq 0$. This implies that $S(t-s,t) \leq s/\eta$. Consequently, since $F \subseteq \overline{\mathbb{R}}_2^+$,

$$III \leq \lambda \int_{t_0}^t \mathbf{P}(B \geq s/\eta; D \geq s)\mathrm{d}s.$$

From this bound, it follows that $III \to 0$ as $t_0 \to \infty$. Since ϑ only has countably many discontinuities, and $S(t-s,t) \to s/z$ on $[0,t_0]$ we have that

$$II \to \lambda \int_0^{t_0} \vartheta(F + (s/z, s))\mathrm{d}s.$$

Taking $t \to \infty$ and then $t_0 \to \infty$ yield the first statement of the theorem.

To prove the second statement, note that (3.6) has a unique solution for $F = \overline{\mathbb{R}}_2^+$, which uniquely determines $S(s,t)$ for all s,t with $s \leq t$. Since ζ is completely determined by ζ_0, and $S(s,t)$, uniqueness follows. □

4. Applications

In this section we analyze a number of quantitative properties of the fluid model equation (2.9). In particular, we investigate the fixed point equation

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

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.

We first give a heuristic interpretation of Equation (4.1): Let Z^r denote the steady-state number of customers in the system. Furthermore, let $V^r(B)$ be the sojourn time of a customer if the customer never reneges. Then the actual sojourn time is given by $\min\{V^r(B), D^r\}$, and from Little's law we get

$$\mathbf{E}\left[Z^r\right] = \lambda \mathbf{E}\left[\min\{V^r(B), D^r\}\right]. \tag{4.2}$$

Divide both sides of (4.2) by r. Since we observe the system in steady state at time 0, the number of customers hardly changes and by the snapshot principle we conclude that $V^r = Z^r B + \mathrm{o}(r)$. Furthermore, we have $D^r = Dr$. Noting that $Z^r/r \to z$ then gives (4.1) after dividing both sides of (4.2) by r and letting $r \to \infty$.

Apart from the mean queue length z, we are also interested in the long term fraction of customers that leave the system successfully. Denote this fraction by P_s. It is clear that

$$P_s = \mathbf{P}\left(D > z_\infty B\right).$$

The following remarkable property, which simply follows from the fixed-point equation (4.1), shows that the performance of the system does not depend on the average of D.

Property 4.1. *Consider two systems numbered 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*

$$z_{2,\infty} = a z_{1,\infty}, \quad P_{s,2} = P_{s,1}.$$

We now proceed by analyzing a number of special cases. In Section 4.1, we assume a strong form of dependence. Section 4.2 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, Section 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 where $D = \Theta B$, with $\Theta > 0$ (independent of B) reflecting the average service rate expected by a customer. In this case, the performance measures can be determined from the equations (recall that $\rho = \alpha \mathbf{E}\left[B\right] > 1$)

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

Some specific examples:

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

— Θ *two-valued.* From the previous example, it is clear that the system can only get some work done if some customers 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 non-negative solution, which is the case if and only if $\rho(1-p) < 1$ (i.e. when the most patient customers cannot saturate the system alone). 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 $P_s = 0$ iff $(1-\rho(1-p))\theta_2 < \rho p \theta 1$. If the reverse inequality holds then all customers of type 2 are being served successfully, i.e. $P_s = (1-p)$.

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

Since P_s does not depend on the mean of Θ, and since the worst-case property of the case of constant Θ, it seems natural to conjecture that the system performance is positively related 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 $x \geq 0$.

Thus, if $\Theta_1 \overset{cvx}{\geq} \Theta_2$, it follows that $z_{2,\infty} \geq z_{1,\infty}$ i.e. less variability in reneging behavior implies a lower service rate. To prove that also $\mathbf{P}(\Theta_1 > z_{1,\infty}) \geq \mathbf{P}(\Theta_2 > z_{2,\infty})$ seems hard without imposing further assumptions.

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

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

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

Recall that $P_s = \mathbf{P}(D \geq z_\infty B)$. Consequently, if B is exponentially distributed, we have the insensitivity (w.r.t. the distribution of D) result $P_s = 1/\rho$. The inequality $P_s \leq 1/\rho$ holds if B^* is stochastically dominated by B, and $P_s \geq 1/\rho$ vice versa. Since B^* being stochastically dominated by B is related to a 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 (and B a general distribution), we see that z_∞ is the solution of

$$\rho\beta^*(z_\infty\nu)) = 1, \tag{4.3}$$

with $\beta^*(s) = \mathbf{E}\left[\mathrm{e}^{-sB^*}\right]$.

In addition, we have the following remarkable expression for the complete fluid limit $z(t), t \geq 0$, if $z_0 = 0$:

Proposition 4.2. *Suppose* $\mathbf{P}(D \geq t) = \mathrm{e}^{-\nu t}$, *that* B *is independent of* D *and that* $z_0 = 0$. *Then the unique solution of (2.9) is given by*

$$z(t) = z_\infty(1 - \mathrm{e}^{-\nu t}), \tag{4.4}$$

with z *the solution of Equation (4.3).*

Proof. Recall that Equation (2.9) has a unique solution. We show that (4.4) is indeed the solution of (2.9) by verification. We thus compute the right hand side of (2.9) writing $z(u) = z_\infty(1 - \mathrm{e}^{-\nu u})$.

Observe that

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

Consequently,

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

Which shows that $z_\infty(1 - \mathrm{e}^{-\nu t})$ satisfies (2.9).

□

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

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

When we view PS as a way of modeling TCP, this example models the integration of elastic (TCP) traffic and TCP friendly 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) for q specializes to

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

Consequently, if the stability condition $\lambda p \mathbf{E}[B_1]$ 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 2.2, that is, we show that the sequence of processes $\{\overline{\mathcal{Z}}^r(\cdot), r \in \mathcal{R}\}$ is tight in $\mathbf{D}([0,\infty), \mathbf{M_1})$. The main results in this section implying this property are the compact containment 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 $\overline{\mathcal{Z}}^r(\cdot)$. The compact containment property is derived in Section 5.3. Section 5.4 serves as a preparation for the oscillation bound. In particular, it is shown that $\overline{\mathcal{Z}}^r(t)$ charges arbitrarily small mass to thin L-shaped sets. The oscillation bound is then shown in Section 5.5.

Throughout this section, it is assumed that the assumptions of Section 2.2 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 convenient to consider them together as a single, measure valued arrival process. For $r \in \mathcal{R}$ and $t \geq s \geq 0$, define the fluid scaled measure valued arrival process by

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

and define the fluid scaled increment

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

Note that $\overline{\mathcal{L}}^r(\cdot)$ is a random element of $\mathbf{D}([0,\infty), \mathbf{M_1})$ and, for each $t \geq s \geq 0$, $\overline{\mathcal{L}}^r(s,t)$ is a random element of $\mathbf{M_1}$.

To state and prove the result, we first introduce some notions from empirical process theory. Our primary reference is [32]. A collection $\mathcal{C}$ 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 $\{C \cap \{x_1, \ldots, x_n\} : C \in \mathcal{C}\}$ has cardinality 2^n. In this case, say that $\mathcal{C}$ *picks out* all subsets of $\{x_1, \ldots, x_n\}$. The *Vapnik-Červonenkis index (VC-index)* of $\mathcal{C}$ is

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

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

VC-classes satisfy a useful entropy bound. Let $\mathcal{Q}$ denote the set of Borel probability measures on $\overline{\mathbb{R}}_+^2$ and, for $Q \in \mathcal{Q}$, 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 $\mathcal{C}$ is a VC-class, then for all $\varepsilon > 0$, the family $\mathscr{V} = \{1_C : C \in \mathcal{C}\}$ satisfies

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

see Theorem 2.6.4 in [32].

Recall the collection of corner sets $\mathcal{C}$ defined in Section 2.3:

$$\mathcal{C} = \{[x,\infty)\times[y,\infty) : x,y\in\mathbb{R}_+\} \cup \left\{[x,\infty]\times[y,\infty] : x,y\in\overline{\mathbb{R}}_+\right\}.$$

Note that for any 3-point subset $\{x_1,x_2,x_3\}\subset\overline{\mathbb{R}}_+^2$, it is impossible for $\mathcal{C}$ to pick out all three 2-point subsets of $\{x_1,x_2,x_3\}$. Since $\mathcal{C}$ shatters no 3-point subset, it has VC-index bounded above by 3. Thus, $\mathcal{C}$ is a VC-class and $\mathscr{V} = \{1_C : C\in\mathcal{C}\}$ 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 $\pi(x,y) = \max\{x,y\}$. Since π is continuous, (2.5) and the Skorohod representation theorem imply the existence of $\overline{\mathbb{R}}_+$-valued random variables $X^r \sim \breve{\vartheta}^r\circ\pi^{-1}$ and $X\sim\vartheta\circ\pi^{-1}$ such that $X^r\to X$ almost surely. Thus, there exists an $\overline{\mathbb{R}}_+$-valued random variable Y such that

$$Y = \sup_{r\in\mathcal{R}} X^r, \quad \text{almost surely.} \tag{5.3}$$

Let μ be the law of Y on $\overline{\mathbb{R}}_+$. Since $L_2(\mu)$ 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 $C\in\mathcal{C}$. 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 $r\to\infty$,*

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

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

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

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. Since $\langle f,\overline{\mathcal{L}}^r(t)\rangle$ 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 $t\in[0,T]$,

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

Since

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

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

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

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

The first equation follows from assumption (2.4) and by observing that

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

which follows from (5.3) and (5.4). To show (5.6), it suffices to verify three assumptions of Theorem 2.8.1 in [32]. Observe that for each $n \in \mathbb{N}$ and $(e_1, \dots, e_n) \in \mathbb{R}^n$, the function

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

is measurable on the completion of $(\overline{\mathbb{R}}_+^2, \mathscr{B}, \breve{\vartheta}^r)^n$, for each $r \in \mathcal{R}$. Thus, $\overline{\mathscr{V}}$ is a $\breve{\vartheta}^r$*-measurable class* for each $r \in \mathcal{R}$; see Definition 2.3.3 in [32]. Moreover, $\overline{\mathscr{V}}$ is uniformly bounded above by the envelope function F, and

$$\lim_{M\to\infty} \sup_{r\in\mathcal{R}} \langle F 1_{\{F>M\}}, \breve{\vartheta}^r \rangle = 0,$$

by Markov's inequality, (5.3), and (5.4). Lastly, $\overline{\mathscr{V}}$ satisfies the finite entropy bound (5.2) because $N(\varepsilon, \overline{\mathscr{V}}, L_1(Q)) \le N(\varepsilon, \mathscr{V}, L_1(Q)) + 1$ and $\mathcal{C}$ is a VC-class. The previous three observations imply that the assumptions of Theorem 2.8.1 in [32] are satisfied. Consequently, $\overline{\mathscr{V}}$ is *Glivenko-Cantelli, uniformly in* r. That is, for every $\delta > 0$, there exists an n_δ such that $n \ge n_\delta$ implies

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

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

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

The first term equals zero by (2.4). For the second term, rewrite

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

and bound each probability in the second term by

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

By (2.4), 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 $r \in \mathcal{R}$. This implies (5.6). □

5.2. **Fluid scaled dynamic equation.** Using (2.3), 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 $t, h \geq 0$,

$$\overline{\mathcal{Z}}^r(t+h)(A) = \overline{\mathcal{Z}}^r(t)\left(A + (\overline{S}^r(t,t+h),h)\right) + \sum_{i=r\overline{E}^r(t)+1}^{r\overline{E}^r(t+h)} 1_A^+\left(\overline{B}_i^r(t+h), \overline{D}_i^r(t+h)\right). \tag{5.10}$$

Subsequent proofs use estimates obtained from this equation. Two estimates result from bounding the summands in (5.10) by 1 and optionally bounding the first term on the right side by its total mass; for each $A \in \mathscr{B}$ and $t, h \geq 0$,

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

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

$$\overline{\mathcal{Z}}^r(t)\left(A + (\overline{S}^r(t,t+h),h)\right) \leq \overline{\mathcal{Z}}^r(t+h)(A) \leq \overline{\mathcal{Z}}^r(t+h)\left(\overline{\mathbb{R}}_+^2\right). \tag{5.12}$$

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 $\mathbf{K} \subset \mathbf{M_1}$ such that*

$$\liminf_{r\to\infty} \mathbf{P}^r\left(\overline{\mathcal{Z}}^r(t) \in \mathbf{K} \text{ for all } t \in [0,T]\right) \geq 1-\eta. \tag{5.13}$$

Proof. A set $\mathbf{K} \subset \mathbf{M_1}$ 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 [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 (2.6), $\overline{\mathcal{Z}}^r(0) \xrightarrow{\mathbf{w}} \zeta_0$ in distribution, and so the sequence $\{\overline{\mathcal{Z}}^r(0)\}$ is tight. Thus, there is a compact set $\mathbf{K}_0 \subset \mathbf{M_1}$, such that

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

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 $n \in \mathbb{N}$. Since $\mathbf{K}_0$ 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 $\lim_{n\to\infty} \sup_{\xi\in\mathbf{K}_0} \xi(J_n^c) = 0$. Since $J_n \subset K_{k(n)}$ 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 Section 5.1 of the envelope function $F = \psi \circ \pi$ for the family $\overline{\mathscr{V}}$. By (2.4) and (5.7), the constant $M = \sup_{r\in\mathcal{R}} \left(\lambda^r T\langle F, \breve{\vartheta}^r\rangle + 1\right)$ is finite. Let $\mathbf{K}$ be the closure of the set

$$\left\{\xi \in \mathbf{M_1} : \xi(\overline{\mathbb{R}}_+^2) \leq M_0 + M \text{ and } \xi(K_n^c) \leq a_n + \psi(n)^{-1}M \text{ for all } n \in \mathbb{N}\right\}.$$

Since $a_n + \psi(n)^{-1}M \to 0$ as $n \to \infty$, the set $\mathbf{K}$ is compact in $\mathbf{M_1}$.

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

$$\Omega_1^r = \left\{ \langle F, \overline{\mathcal{L}}^r(T) \rangle \leq \lambda^r T \langle F, \breve{\vartheta}^r \rangle + 1 \right\}.$$

By (5.14) and Lemma 5.1, $\liminf_{r\to\infty} \mathbf{P}^r(\Omega_0^r \cap \Omega_1^r) \geq 1 - \eta$. 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 $\overline{\mathcal{Z}}^r(t) \in \mathbf{K}$.

By (5.11),

$$\overline{\mathcal{Z}}^r(t)(\overline{\mathbb{R}}_+^2) \leq \overline{\mathcal{Z}}^r(0)(\overline{\mathbb{R}}_+^2) + \overline{\mathcal{L}}^r(t)(\overline{\mathbb{R}}_+^2).$$

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

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

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

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

The shape of the set K_n^c implies that

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

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

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

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

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

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

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

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

5.4. **Asymptotic regularity.** The second and main step necessary to prove tightness is to bound the probability that the process $\overline{\mathcal{Z}}^r(\cdot)$ 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 $\overline{\mathcal{Z}}^r(\cdot)$ assigns arbitrarily small mass to the boundaries of the sets $C \in \mathcal{C}$. This is phrased in terms of κ-enlargements of the boundaries of these sets (forming a collection of L-shaped sets). For $C \in \mathcal{C}$ and $\kappa > 0$, let ∂_C denote the boundary of C and let

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

be the κ-enlargement in $\overline{\mathbb{R}}_+^2$ of its boundary, where the infimum over the empty set equals ∞. (Note that ∂_C, and therefore also ∂_C^κ, is empty for the corner sets $\overline{\mathbb{R}}_+^2$ and $\{\infty\} \times \{\infty\}$. 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 $\overline{\mathcal{Z}}^r(0)$.

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

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

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

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

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

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

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

The projection $(x, y) \mapsto x$ is continuous, so (2.6) implies that $\overline{\mathcal{Z}}_1^r(0)$ converges in distribution to $\zeta_0(\cdot \times \overline{\mathbb{R}}_+)$ as $r \to \infty$. Since $\zeta_0(\cdot \times \overline{\mathbb{R}}_+)$ is free of atoms in $[0,\infty)$, there exists a $\kappa > 0$ such that

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

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

$$\zeta_0([M,\infty) \times \overline{\mathbb{R}}_+) \le \varepsilon. \tag{5.20}$$

Let $N = \lceil M/\kappa \rceil + 1$, where $\lceil x \rceil$ denotes the smallest integer $n \ge x$. For $n = 1, \ldots, N-1$, 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 $n \le N$ such that $[x, x+2\kappa] \subset I_n$. To prove (5.18), it therefore suffices to show that

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

Let $\mathbf{M_1}(\overline{\mathbb{R}}_+)$ denote the space of finite nonnegative Borel measures on $\overline{\mathbb{R}}_+$, endowed with the weak topology. Let $\mathbf{A} = \{\xi \in \mathbf{M_1}(\overline{\mathbb{R}}_+) : \max_{n\le N} \xi(I_n) < \varepsilon\}$, and suppose that a sequence $\{\xi_k\} \subset \mathbf{M_1}(\overline{\mathbb{R}}_+)$ satisfies $\xi_k \xrightarrow{\mathbf{w}} \xi$ for some $\xi \in \mathbf{A}$. Since the sets I_n are closed, the Portmanteau theorem (adapted to finite measures) implies that

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

Hence, $\xi_k \in \mathbf{A}$ for sufficiently large k, which implies that $\mathbf{A}$ is open in $\mathbf{M_1}(\overline{\mathbb{R}}_+)$. Thus, a second application of the Portmanteau theorem yields

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

which implies (5.21). □

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

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

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

Proof. By Lemmas 5.1, 5.2, and 5.3, there exists a compact $\mathbf{K} \subset \mathbf{M_1}$ and a $\kappa_0 > 0$, such that for all $\delta > 0$, the events

$$\begin{aligned}
\Omega_1^r &= \left\{ \sup_{C \in \mathcal{C}} \overline{\mathcal{Z}}^r(0)(\partial_C^{\kappa_0}) \le \frac{\varepsilon}{2} \right\}, \\
\Omega_2^r &= \left\{ \sup_{C \in \mathcal{C}} \sup_{0 \le s \le t \le T} \left| \overline{\mathcal{L}}^r(s,t)(C) - \lambda^r(t-s)\breve{\vartheta}^r(C) \right| \le \delta \right\}, \\
\Omega_3^r &= \left\{ \overline{\mathcal{Z}}^r(t) \in \mathbf{K} \text{ for all } t \in [0,T] \right\}, \\
\Omega_0^r &= \Omega_1^r \cap \Omega_2^r \cap \Omega_3^r,
\end{aligned}$$

satisfy

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

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

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

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

Let $\lambda^* = \sup_{r \in \mathcal{R}} \lambda^r$, which is finite by (2.4). Fix

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

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

$$\tau_1 = \sup\{s \le t : \langle 1, \overline{\mathcal{Z}}^r(s)\rangle = 0\},$$

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

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

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

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

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

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

Letting $\tau_n \uparrow \tau$ yields

$$\overline{\mathcal{Z}}^r(\tau)\left(\partial_C^\kappa + (\overline{S}^r(\tau,t), t-\tau)\right) \le \delta \le \frac{\varepsilon}{2}.$$

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

$$\overline{S}^r(\tau, t) = \int_{t-RM}^{t} \langle 1, \overline{\mathcal{Z}}^r(s)\rangle^{-1} \, \mathrm{d}s \geq R.$$

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

$$\overline{\mathcal{Z}}^r(\tau)\left(\partial_C^\kappa + (\overline{S}^r(\tau, t), t - \tau)\right) \leq \overline{\mathcal{Z}}^r(\tau)\,(K_R^c) \leq \frac{\varepsilon}{2},$$

which proves (5.26).

By (5.10),

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

Let I denote the second right hand term in (5.27). By (5.26), it remains to show that $I \leq \varepsilon/2$. 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 $1^+_{\partial_C^\kappa}(\cdot,\cdot) \leq 1_{\partial_C^\kappa}(\cdot,\cdot)$,

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

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

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

By definition,

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

So, letting

$$\begin{aligned} C_n^- &= C + \left(\overline{S}^r(t^n, t) - \kappa, t - t^n - \kappa\right) \cap \overline{\mathbb{R}}_2^+, \\ C_n^+ &= C + \left(\overline{S}^r(t_n, t) + \kappa, t - t_n + \kappa\right) \cap \overline{\mathbb{R}}_2^+, \\ C_n &= C_n^- \setminus C_n^+, \end{aligned}$$

it follows from (5.29) and (5.30) that

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

Conclude from (5.28) and (5.31) that

$$I \leq \sum_{n=0}^{N-1} \frac{1}{r} \sum_{i=r\overline{E}^r(t_n)+1}^{r\overline{E}^r(t^n)} 1_{C_n}(B_i^r, D_i^r r^{-1}) = \sum_{n=0}^{N-1} \left(\overline{\mathcal{L}}^r(t_n, t^n)(C_n^-) - \overline{\mathcal{L}}^r(t_n, t^n)(C_n^+)\right).$$

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

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

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

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

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

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

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

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

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

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

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 $\mathbf{M_1}$, we use the Prohorov metric (adapted to finite measures). For $\mu, \nu \in \mathbf{M_1}$, define

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

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

Definition 5.5. *For each $\zeta(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M_1})$ 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*

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

Proof. As before, let $\lambda^* = \sup_{r\in\mathcal{R}} \lambda^r$. 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

$$\Omega_1^r = \left\{ \sup_{t\in[0,T]} \overline{\mathcal{Z}}^r(t)(L_\kappa) \le \frac{\varepsilon}{4} \right\},$$
$$\Omega_2^r = \left\{ \sup_{t\in[0,T-\delta]} \overline{\mathcal{L}}^r(t,t+\delta)(\overline{\mathbb{R}}_+^2) \le 2\lambda^*\delta \right\},$$
$$\Omega_0^r = \Omega_1^r \cap \Omega_2^r,$$

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 $\Omega_0^r \subset \Omega_*^r$ for each r. Fix $r \in \mathcal{R}$ and $\omega \in \Omega_0^r$; 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 $A \in \mathscr{B}$ be closed. It suffices to show the two inequalities,

$$\overline{\mathcal{Z}}^r(t)(A) \le \overline{\mathcal{Z}}^r(t+h)(A^\varepsilon) + \varepsilon, \tag{5.35}$$

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

To show (5.35), use the definition of Ω_1^r to write

$$\begin{aligned} \overline{\mathcal{Z}}^r(t)(A) &\le \overline{\mathcal{Z}}^r(t)(L_\kappa) + \overline{\mathcal{Z}}^r(t)\left(A \cap L_\kappa^c\right) \\ &\le \frac{\varepsilon}{4} + \overline{\mathcal{Z}}^r(t)\left(A \cap L_\kappa^c\right). \end{aligned} \tag{5.37}$$

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

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

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

$$A \cap L_\kappa^c \subset A^\varepsilon + (\overline{S}^r(t,t+h),h). \tag{5.39}$$

Deduce from (5.37) that

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

Apply (5.12) to get

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

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

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

By (5.37) and (5.41),

$$\overline{\mathcal{Z}}^r(t)(A) \le \frac{\varepsilon}{4} + \overline{\mathcal{Z}}^r(t)(L_\kappa^c) \le \frac{\varepsilon}{4} + \overline{\mathcal{Z}}^r(t)\left(\overline{\mathbb{R}}_+^2 + (\overline{S}^r(t,\tau),\tau-t)\right).$$

Apply (5.12) to get

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

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

To show (5.36), use (5.11) and the definitions of Ω_2^r and δ to obtain

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

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

$$\overline{\mathcal{Z}}^r(t+h)(A) \leq \overline{\mathcal{Z}}^r(t)\left(A^\varepsilon\right) + \frac{\varepsilon}{4}.$$

If $I \neq \emptyset$, then by (5.11), the definition of Ω_2^r and the choice of δ,

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

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

$$\mathbf{d}\left[\overline{\mathcal{Z}}^r(t), \overline{\mathcal{Z}}^r(t+h)\right] \leq \varepsilon.$$

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

$$\mathbf{w}_T\left(\overline{\mathcal{Z}}^r(\cdot),\delta\right) \leq \varepsilon,$$

which implies that $\omega \in \Omega_*^r$. □

6. Limiting Fluid Equations

This section contains the proof of Theorem 2.2. Tightness of the sequence $\{\overline{\mathcal{Z}}^r(\cdot)\}$ follows immediately from Lemmas 5.2 and 5.6. Since $\{\overline{\mathcal{Z}}^r(\cdot)\}$ is tight, there exists a subsequence $\{q\} \subset \mathcal{R}$ and a process $\mathcal{Z}(\cdot)$ in $\mathbf{D}([0,\infty),\mathbf{M})$ such that $\overline{\mathcal{Z}}^q(\cdot) \Rightarrow \mathcal{Z}(\cdot)$ as $q \to \infty$. We must show that $\mathcal{Z}(\cdot)$ is almost surely a measure valued fluid model solution for the data $(\lambda,\vartheta,\zeta_0)$. This is accomplished by Lemmas 6.1 and 6.2, and Theorem 6.3 below. Finally, if (2.10) holds, then a measure valued fluid model solution for $(\lambda,\vartheta,\zeta_0)$ is unique by Theorem 2.3. In this case, the law of the limit point $\mathcal{Z}(\cdot)$ is unique and so $\overline{\mathcal{Z}}^r(\cdot) \Rightarrow \mathcal{Z}(\cdot)$ as $r \to \infty$.

Let $Z(\cdot) = \langle 1, \mathcal{Z}(\cdot)\rangle$ be the total mass process for $\mathcal{Z}(\cdot)$, and let $S(u,v) = \int_u^v \frac{1}{Z(s)}\,\mathrm{d}s$ for all $v \geq u \geq 0$. To show that $\mathcal{Z}(\cdot)$ is almost surely a measure valued fluid model solution, note first that $\mathcal{Z}(\cdot)$ is almost surely continuous by Lemma 5.6. Note also that, by (2.6), $\mathcal{Z}(0) = \zeta_0$ almost surely. It remains to show that properties (i) and (ii) of Definition 2.1 are satisfied almost surely by $\mathcal{Z}(\cdot)$. The next result establishes (i).

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

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

Proof. Take $t > 0$. Pick a constant $a < t$ small enough such that the marginal distribution of D is continuous at a, take $m < \infty$ such that the marginal distribution of B is continuous at m, and such that $\lambda\mathbf{E}\left[B1_{\{D>a,B<m\}}\right] > 1$. By dominated convergence,

$$\lim_{r\to\infty} \lambda^r \mathbf{E}\left[B_1^r 1_{\{D_1^r r^{-1} > a; B_1^r \leq m\}}\right] = \lambda\mathbf{E}\left[B1_{\{D>a;B\leq m\}}\right].$$

Compare the original system with an ordinary PS queues having arrival rate $\lambda^r_{a,m} = \lambda^r \mathbf{P}\left(D^r > ra; B^r < m\right)$ and service times $B^r_{i,a,m}$, which are distributed as $B^r_i \mid D^r_i > ra; B^r_i \leq m$. Suppose that this PS queue is empty at time $r(t-a)$, and let $\acute{Z}^r(t)$ be the queue length in this PS queue at time ra.

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

Since the service times in our modified system are bounded, the means converge. The assumptions in [28] are therefore valid, and it follows that there exists a constant $k_a > 0$ such that $\lim_{r\to\infty} \acute{Z}^r(rt)/r = k_a$ almost surely. Consequently, we have $\liminf_{r\to\infty} \overline{Z}^r(t) \geq m_a$ almost surely, which implies the assertion. □

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

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

$$\mathcal{Z}(t)(\partial_C) = 0. \tag{6.2}$$

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,

$$\liminf_{q\to\infty} \mathbf{P}^q \left(\sup_{t\in[0,T]} \sup_{C\in\mathcal{C}} \overline{\mathcal{Z}}^q(t)(\partial_C^{\kappa_n}) \leq \frac{1}{n} \right) \geq 1 - \eta_n. \tag{6.3}$$

For each $n \in \mathbb{N}$, let $\mathbf{M}_n = \{\xi \in \mathbf{M} : \sup_{C\in\mathcal{C}} \xi(\partial_C^{\kappa_n}) \leq 1/n\}$. 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

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

Thus, $\xi \in \mathbf{M}_n$ and $\mathbf{M}_n$ is closed. By definition of the Skorohod J_1-topology, the set $\mathbf{D}_n^T = \{\zeta(\cdot) \in \mathbf{D}\left([0,\infty), \mathbf{M}\right) : \zeta(t) \in \mathbf{M}_n \text{ for all } t \in [0,T]\}$ is also closed. Apply the Portmanteau theorem and (6.3) to obtain

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

By the Borel-Cantelli lemma,

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

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

$$\sup_{t\in[0,T]} \sup_{C\in\mathcal{C}} \mathcal{Z}(t)(\partial_C^{\kappa_n}) \leq \frac{1}{n}, \quad \text{for all } n > N. \tag{6.4}$$

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

$$\sup_{t\in[0,T]} \sup_{C\in\mathcal{C}} \mathcal{Z}(t)(\partial_C) = 0.$$

□

We now establish property (ii). Recall that $Z(t) = \langle 1, \mathcal{Z}(t)\rangle$ 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 $\mathcal{Z}(\cdot)$ satisfies*

$$\mathcal{Z}(t)\,(A) = \mathcal{Z}(0)\,(A + (S(0,t),t)) + \lambda \int_0^t \vartheta\,(A + (S(s,t), t-s))\,\mathrm{d}s, \tag{6.5}$$

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

Proof. Let $T > 0$. It suffices to show that almost surely, (6.5) holds for all $t \in [0,T]$ and all $A \in \mathscr{B}$. For each $r \in \mathcal{R}$, define the random variable

$$X_T^r = \sup_{C\in\mathcal{C}} \sup_{0\leq s\leq t\leq T} \left|\overline{\mathcal{L}}^r(s,t)(C) - \lambda^r(t-s)\breve{\vartheta}^r(C)\right|. \tag{6.6}$$

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

$$(\overline{\mathcal{Z}}^q(\cdot), X_T^q) \to (\mathcal{Z}(\cdot), 0), \quad \text{almost surely.} \tag{6.7}$$

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), (6.2) and (6.7) hold.

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

$$\overline{\mathcal{Z}}^q(t)(C) = \overline{\mathcal{Z}}^q(0)\left(C + (\overline{S}^q(0,t),t)\right) + \frac{1}{q}\sum_{i=1}^{q\overline{E}^q(t)} 1_C^+\left(\overline{B}_i^q(t), \overline{D}_i^q(t)\right). \tag{6.8}$$

We will obtain (6.5) from (6.8) by letting $q \to \infty$. The convergence in the first component of (6.7) is in the Skorohod J_1-topology on $\mathbf{D}([0,\infty), \mathbf{M_1})$. However, since $\mathcal{Z}(\cdot)$ is continuous,

$$\overline{\mathcal{Z}}^q(s) \xrightarrow{\mathbf{w}} \mathcal{Z}(s), \quad \text{for all } s \in [0,t]. \tag{6.9}$$

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

$$\lim_{q\to\infty} \left\|\overline{Z}^q(\cdot) - Z(\cdot)\right\|_t = 0. \tag{6.10}$$

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

$$\begin{aligned}\lim_{q\to\infty} \overline{S}^q(u,v) &= \lim_{q\to\infty} \int_u^v \frac{1}{\overline{Z}^q(s)}\,\mathrm{d}s \\ &= \int_u^v \frac{1}{Z(s)}\,\mathrm{d}s \\ &= S(u,v).\end{aligned} \tag{6.11}$$

If $Z(0) \neq 0$, then (6.11) 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 $\overline{S}^q(0, v) \to \infty$ as $q \to \infty$.

Suppose that $Z(0) \neq 0$ and let $\varepsilon > 0$. By (6.11), there exists a $q_\varepsilon \in \mathcal{R}$ such that $\overline{S}^q(0,t) \in ((\overline{S}(0,t) - \varepsilon)^+, \overline{S}(0,t) + \varepsilon)$ for $q > q_\varepsilon$. Deduce from the shape of the set C, (6.9), and (6.2) that

$$\limsup_{q\to\infty} \overline{\mathcal{Z}}^q(0) \left(C + (\overline{S}^q(0,t), t)\right) \leq \overline{\mathcal{Z}}(0) \left(C + ((\overline{S}(0,t) - \varepsilon)^+, t)\right),$$
$$\liminf_{q\to\infty} \overline{\mathcal{Z}}^q(0) \left(C + (\overline{S}^q(0,t), t)\right) \geq \overline{\mathcal{Z}}(0) \left(C + (\overline{S}(0,t) + \varepsilon, t))\right).$$

By (6.2), letting $\varepsilon \to 0$ yields

$$\lim_{q\to\infty} \overline{\mathcal{Z}}^q(0) \left(C + (\overline{S}^q(0,t), t)\right) = \overline{\mathcal{Z}}(0) \left(C + (\overline{S}(0,t), t)\right). \tag{6.12}$$

If $Z(0) = 0$, then (6.12) holds trivially because the left side is bounded above by $\lim_{q\to\infty} \langle 1, \overline{\mathcal{Z}}^q(0)\rangle = 0$ by (6.10). Combining with (6.9) and (6.2) for $\overline{\mathcal{Z}}^q(t)$, implies that, as $q \to \infty$,

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

Let I^q denote the second right hand term in (6.8). Let $\delta > 0$ and let $\eta \in (0, t)$. Since $\overline{S}^q(s,t)$ is decreasing in s and $S(\cdot, t)$ is continuous on $[\eta, t]$, (6.11) implies that $\overline{S}^q(\cdot, t) \to S(\cdot, t)$ uniformly on $[\eta, t]$. That is, there exists $q_\delta \in \mathcal{R}$ such that

$$\sup_{s\in[\eta,t]} \left|\overline{S}^q(s,t) - S(s,t)\right| \leq \delta, \qquad \text{for all } q > q_\delta. \tag{6.13}$$

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

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

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 $\max_{j\leq N-1}(t_{j+1}^N - t_j^N) \to 0$ as $N \to \infty$. Then

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

Note that the first right hand term is bounded above by $\overline{\mathcal{L}}^q(0, \eta)(\overline{\mathbb{R}}_+^2)$. Suppose that $t_j^N \leq U_i^q q^{-1} \leq t_{j+1}^N$, for some $q > q_\delta$, some $j \leq N - 1$, and some $i \in \{q\overline{E}^q(\eta) + 1, \ldots, q\overline{E}^q(t)\}$. Then by (6.13),

$$S(t_{j+1}^N, t) - \delta \leq \overline{S}^q(U_i^q q^{-1}, t) \leq S(t_j^N, t) + \delta. \tag{6.14}$$

By definition,

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

So for $q > q_\delta$, (6.14) and the inequalities $1_C(\cdot - \delta, \cdot) \le 1_C^+(\cdot, \cdot) \le 1_C(\cdot + \delta, \cdot)$ yield

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

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

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

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

Rewrite as

$$\begin{aligned} I^q &\ge \sum_{j=0}^{N-1} \overline{\mathcal{L}}^q(t_j^N, t_{j+1}^N)\left(C + \left(S(t_j^N, t) + 2\delta, t - t_j^N\right)\right); \\ I^q &\le \overline{\mathcal{L}}^q(0, \eta)(\overline{\mathbb{R}}_+^2) + \sum_{j=0}^{N-1} \overline{\mathcal{L}}^q(t_j^N, t_{j+1}^N)\left(C + \left(S(t_{j+1}^N, t) - 2\delta, t - t_{j+1}^N\right)\right). \end{aligned} \tag{6.15}$$

By (6.6) and (6.15), $q > q_\delta$ implies that

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

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

By (6.7), and since $t_j^N \notin D_\vartheta(S)$ for all $j = 1, \dots, N-1$,

$$\begin{aligned} \liminf_{q\to\infty} I^q &\ge \lambda \sum_{j=0}^{N-1} (t_{j+1}^N - t_j^N)\vartheta\left(C + \left(S(t_j^N, t) + 2\delta, t - t_j^N\right)\right); \\ \limsup_{q\to\infty} I^q &\le \lambda\eta + \lambda \sum_{j=0}^{N-1} (t_{j+1}^N - t_j^N)\vartheta\left(C + \left(S(t_{j+1}^N, t) - 2\delta, t - t_{j+1}^N\right)\right). \end{aligned} \tag{6.16}$$

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

$$\begin{aligned} &\lim_{N\to+\infty} \sum_{j=0}^{N-1} 1_{[t_j^N, t_{j+1}^N)}(s)\vartheta\left(C + \left(S(t_j^N, t) + 2\delta, t - t_j^N\right)\right) \\ &\qquad = \vartheta\left(C + (S(s, t) + 2\delta, t - s)\right); \\ &\lim_{N\to+\infty} \sum_{j=0}^{N-1} 1_{[t_j^N, t_{j+1}^N)}(s)\vartheta\left(C + \left(S(t_{j+1}^N, t) - 2\delta, t - t_{j+1}^N\right)\right) \\ &\qquad = \vartheta\left(C + (S(s, t) - 2\delta, t - s)\right). \end{aligned} \tag{6.17}$$

Thus, the convergence in (6.17) holds for almost every $s \in [\eta, t)$. Let $N \to \infty$ in (6.16) and conclude from (6.17) and the bounded convergence theorem that

$$\begin{aligned} \liminf_{q\to\infty} I^q &\geq \lambda \int_\eta^t \vartheta\left(C + (S(s,t) + 2\delta, t-s)\right) \mathrm{d}s; \\ \limsup_{q\to\infty} I^q &\leq \lambda\eta + \lambda \int_\eta^t \vartheta\left(C + (S(s,t) - 2\delta, t-s)\right) \mathrm{d}s. \end{aligned} \tag{6.18}$$

Let $\delta \to 0$ in (6.18). Since $D_\vartheta(\mathcal{C})$ is countable, both integrands in (6.18) converge almost everywhere on $[\eta, t]$ to $\vartheta\left(C + (S(s,t), t-s)\right)$. Thus,

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

Let $\eta \to 0$ to conclude that

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

This proves (6.5) for all $t \in [0, T]$ and $C \in \mathcal{C}$. To extend to all $A \in \mathscr{B}$, let $\mathcal{C}'$ be the set of $A \in \mathscr{B}$ for which (6.5) holds. Observe that $\mathcal{C}'$ is a λ-system: $\overline{\mathbb{R}}_+^2 \in \mathcal{C}'$ because $\overline{\mathbb{R}}_+^2 \in \mathcal{C}$; if $\{A_n\} \subset \mathcal{C}'$ satisfies $A_n \uparrow A$, then $A \in \mathcal{C}'$; if $A_1 \subset A_2$ are elements of $\mathcal{C}'$, then $A_2 \setminus A_1 \in \mathcal{C}'$. Observe also that $\mathcal{C}$ is a π-system: if $C_1, C_2 \in \mathcal{C}$, then $C_1 \cap C_2 \in \mathcal{C}$. Since $\mathcal{C} \subset \mathcal{C}'$ and the σ-algebra generated by $\mathcal{C}$ is equal to $\mathscr{B}$, it follows that $\mathcal{C}' = \mathscr{B}$ by the Dynkin $\pi\lambda$-theorem (see for example [2]). □

References

[1] D. Barrer, *Queueing with impatient customers and ordered service*, Operations Research **5** (1957), 650–656.

[2] Patrick Billingsley, *Probability and measure*, 2 ed., John Wiley & Sons, Inc. New York, 1986.

[3] Thomas Bonald and Laurent Massoulié, *Impact of fairness on Internet performance*, Proceedings of ACM Sigmetrics 2001, 2001, pp. 82–91.

[4] Thomas Bonald and James Roberts, *Congestion at flow level and the impact of user behaviour*, Computer Networks **42** (2003), 521–536.

[5] N. Boots and Tijms H., *A multi-server queueuing system with impatient customers*, Management Science **45** (1999), 444–448.

[6] M. Bramson, *Stability of networks for max-min fair routing*, Presentation at the 13th INFORMS Applied Probability Conference, Ottawa, 2005.

[7] E. Coffman, A. Puhalskii, M. Reiman, and P. Wright, *Processor shared buffers with reneging*, Performance Evaluation **19** (1994), 25–46.

[8] Gustavo de Veciana, Takis Konstantopoulos, and Tae-Jin Lee, *Stability and performance analysis of networks supporting elastic services*, IEEE/ACM Trans. Netw. **9** (2001), no. 1, 2–14.

[9] Bogdan Doytchinov, John Lehoczky, and Steven Shreve, *Real-time queues in heavy traffic with earliest-deadline-first queue discipline*, Annals of Applied Probability **11** (2001), no. 2, 332–378.

[10] N. Gans, G. Koole, and A. Mandelbaum, *Telephone call centers: Tutorial, review, and research prospects*, Manufacturing & Service Operations Management **5** (2002), 79–141.

[11] Christian Gromoll, Philippe Robert, Bert Zwart, and Richard Bakker, *The impact of reneging in processor sharing queues*, ACM-Sigmetrics (Saint Malo), ACM/IFIP WG 7.3, June 2006.

[12] H. C. Gromoll and Ł. Kruk, *Heavy traffic analysis of a real-time processor sharing queue*, to appear, 2006.

[13] H. C. Gromoll and R. J. Williams, *Fluid approximation for an Internet congestion control model with fair bandwidth sharing and general document size distributions*, Preprint, 2006.
[14] Fabrice Guillemin, Philippe Robert, and Bert Zwart, *Tail asymptotics for processor-sharing queues*, Advances in Applied Probability **36** (2004), 525–543.
[15] J. Hale and S. Verduyn Lunel, *An introduction to functional differential equations*, Springer Verlag, New York, 1993.
[16] Alain Jean-Marie and Philippe Robert, *On the transient behavior of some single server queues*, Queueing Systems, Theory and Applications **17** (1994), 129–136.
[17] Olav Kallenberg, *Random Measures*, Academic Press, New York, 1986.
[18] F. P. Kelly and R. J. Williams, *Fluid model for a network operating under a fair bandwidth sharing policy*, Annals of Applied Probability **14** (2004), 1055–1083.
[19] P. Key, L. Massoulié, A. Bain, and F. Kelly, *Fair internet traffic integration: Network flow models and analysis*, Annals of Telecommunications **59** (2004), 1338–1352.
[20] L. Kruk, J. Lehoczky, and S. Shreve, *Second order approximation for the customer time in queue distribution under the FIFO service discipline*, Annales UMCS Informatica AI **1** (2003), 37–48.
[21] ———, *Accuracy of state space collapse for earliest-deadline-first queues*, Preprint, 2004.
[22] L. Kruk, J. Lehoczky, S. Shreve, and S. Yeung, *Multiple-input heavy-traffic real-time queues*, Annals of Applied Probability **13** (2003), no. 1, 54–99.
[23] ———, *Earliest-deadline-first service in heavy-traffic acyclic networks*, Annals of Applied Probability **14** (2004), no. 3, 1306–1352.
[24] A. Lakshmikantha, C. L. Beck, and R. Srikant, *Connection level stability analysis of the internet using the sum of squares (SoS) techniques*, Conference on Information Sciences and Systems, Princeton, 2004.
[25] Laurent Massoulié, *Structural properties of proportional fairness: stability and insensitivity*, Preprint, 2005.
[26] Laurent Massoulié and James Roberts, *Bandwidth sharing: Objectives and algorithms*, INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, 1999, pp. 1395–1403.
[27] J. Mo and J. Walrand, *Fair end-to-end window-based congestion control*, IEEE/ACM Transactions on Networking **8** (2000), no. 5, 556–567.
[28] A. L. Puha, A. L. Stolyar, and R. J. Williams, *The fluid limit of an overloaded processor sharing queue*, Preprint, 2004.
[29] James Roberts and Laurent Massoulié, *Bandwidth sharing and admission control for elastic traffic*, Telecommunication Systems **15** (2000), 185–201.
[30] Robert E. Stanford, *Reneging phenomena in single channel queues*, Mathematics of Operations Research **4** (1979), 162–178.
[31] ———, *On queues with impatience*, Advances in Applied Probability **22** (1990), no. 3, 768–769.
[32] Aad van der Vaart and Jon A. Wellner, *Weak convergence and empirical processes*, Springer-Verlag, New York, 1996.
[33] A. Ward and P. Glynn, *A diffusion approximation for a markovian queue with reneging*, Queueing Systems **43** (2003), 103–128.
[34] Shu-Ngai Yeung and John P. Lehoczky, *Real-time queueing networks in heavy traffic with EDF and FIFO queue discipline*, Preprint, 2004.

(Christian Gromoll) Department of Mathematics, Stanford University 450 Serra Mall, Stanford, CA 94305-2125, USA
E-mail address: gromoll@math.stanford.edu

(Ph. Robert) INRIA-Rocquencourt, RAP project, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France
E-mail address: Philippe.Robert@inria.fr
URL: http://www-rocq.inria.fr/~robert

(Bert Zwart) Eindhoven University of Technology, Department of Mathematics and Computer Science, HG 9.35, P.O. Box 513, 5600 MB Eindhoven, the Netherlands
E-mail address: zwart@win.tue.nl