Law of large numbers for the many-server earliest-deadline-first queue

Rami Atar* Anup Biswas† Haya Kaspi‡

September 14, 2017

Abstract

A many-server queue operating under the earliest deadline first discipline, where the distributions of service time and deadline are generic, is studied at the law of large numbers scale. Fluid model equations, formulated in terms of the many-server transport equation and the recently introduced measure-valued Skorohod map, are proposed as a means of characterizing the limit. The main results are the uniqueness of solutions to these equations, and the law of large numbers scale convergence to the solutions.

AMS subject classifications: 60F17, 60G57, 68M20

Keywords: measure-valued processes, measure-valued Skorohod map, many-server transport equation, fluid limits, earliest-deadline-first, least-patient-first, many-server queues

1 Introduction

This paper proves a law of large numbers (LLN) many-server limit for a queueing model with general service time and deadline distributions, operating under the *earliest-deadline-first* (EDF) scheduling policy (we refer to this model as G/G/N+G EDF). By a *many-server limit* we refer to a setting where the number of servers grows without bound. The limit is characterized in terms of a set of so-called *fluid model equations* (FME) that involve both the *many-server transport equation* (MSTE) [15] and the recently introduced *measure-valued Skorohod map* (MVSM) [2]. It provides the first result on the EDF policy involving a many-server limit. Several papers have analyzed EDF asymptotically by appealing to the so-called *frontier process* (see below). However, as argued in [2], the method based on this process is not generic enough to cover a large variety of models (especially ones with time-varying parameters). Our motivation is to extend the asymptotic analysis of EDF to settings where the method involving frontier process is not expected to be effective; the many server regime offers a natural setting of this sort (even when the parameters are constant over time).

*Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel.

†Department of Mathematics, Indian Institute of Science Education and Research, Pune 411008, India.

‡Department of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel.

In recent years the use of measure-valued processes in mathematical modeling of queueing systems has been very successful. As far as many-server asymptotics are concerned, it is well understood since the work of Halfin and Whitt [11] that exponential service time distribution leads to simple limit dynamics; specifically, the diffusion-scale heavy traffic limit of [11] is characterized by a diffusion process on the real line. However, there is a great deal of motivation to study many-server models under general service time distributions, stemming from applications such as call centers and cloud computing. For example, the statistical study of a call center by Brown et al. [6] finds a good fit of the service time data to the lognormal distribution. In this vein, in [24], Whitt considered a G/G/N system with abandonment and proposed a deterministic LLN (otherwise referred to as fluid) approximation. In [15], Kaspi and Ramanan obtained measure-space valued fluid limits for such systems. Kang and Ramanan generalized this work by modeling customer abandonment [13]. Further generalizations in this direction were obtained by Zhang [25] and Walsh-Zuñiga [23]. In [4], Atar et al. studied multi-class many-server queues with fixed priority and established the existence of a unique fluid limit. Kang and Ramanan studied ergodic properties of the G/G/N+G model and its relation with the invariant states of the fluid limit in [14]. Reed [22] established the fluid and diffusion limits of the customer-count processes for many-server queueing systems under a finite first moment assumption on service time distribution.

The aforementioned works were concerned with the *first-come first-served* (FCFS) discipline. Many-server systems operating under the EDF discipline were considered recently by Mandelbaum and Momćilović [18], where a fluid limit heuristic was developed. Motivated by prioritizing customers with least patience and emphasizing the importance of this policy for emergency services, [18] refers to this policy as *least-patience-first*. Decreusefond and Moyal [7] study the fluid limits of M/M/1+M EDF, and Atar et al. [3] generalize these results to G/G/1+G EDF. Diffusion limits for G/G/1+G EDF systems were studied by Doytchinov et al. [9] and Kruk et al. [17]. For additional work on EDF in asymptotic regimes other than the many-server limit we refer to Kruk [16] and references therein.

An attractive feature of EDF, established in several of the aforementioned settings, is that it minimizes the abandonment count. Specifically, in [19, 20, 21] it was shown for a single server model that EDF minimizes customer abandonments within a certain class of scheduling policies. The paper [17] studies G/G/1+G and shows that the *reneged work* is minimized under EDF. Comments in the introduction of [18] also address this minimality property, and so do some of the results of Section 4.1.5 of [2].

In this article we are interested in the many-server LLN limit of the G/G/N+G EDF. To elaborate on the hurdles in obtaining fluid limits in this setting let us briefly mention some tools that have been used in the literature to treat the LLN for the single server EDF. The aforementioned frontier process has been one of the main tools in [3, 9, 17]. One defines the lead time of a customer at time t as the (possibly negative) difference between the customer's deadline and the time t. The frontier process at time t is defined as the maximum lead time at t of all the customers that have ever been in service in the interval $[0, t]$. The method developed in the papers mentioned above relies on the validity of the property that, in an asymptotic sense, the frontier at any given time separates the population of customers to those that have been sent to service and those that are still in the buffer, according as their relative deadline at that time is below the value of the frontier or above it, respectively. The main idea of using the frontier process asymptotics to characterize the asymptotics of the full model is that when this property

is valid, the frontier can be recovered directly from the model's primitive data (specifically, in terms of a one-dimensional Skorohod map), and at the same time the full state of the system, including the measure-valued queueing process, can be expressed in terms of the frontier.

It was argued in [2] that this method breaks down for more general settings than those considered in the above papers, because the property by which the two populations are asymptotically separated by the frontier, fails to hold. Such settings include the single-server model at fluid scale with time varying characteristics (such as rate of arrivals, service or patience distribution). An alternative approach developed in [2] was to introduce a certain Skorohod-type transformation in the space of paths taking values in the space of finite measures on the real line, referred to as a measure-valued Skorohod map (MVSM) (see Proposition 2.1). It was used there to obtain fluid limits of several queueing models, including the G/G/1+G EDF with (possibly) time varying characteristics.

Consider now the many-server asymptotics of the G/G/N+G system. As far as the FCFS discipline is concerned, it is well understood, and captured by the MSTE of [15], that the time evolution of the ages of jobs in the service pool affects the rate at which jobs are transferred from the buffer to the pool. In particular, this rate varies over time even for a model with time homogeneous characteristics, unless the system is at equilibrium (or when the service time distribution is exponential). Now, the time evolution of the content of the buffer under variable rate of transferring jobs to the service pool is much like that of a single server model for which the service rate varies over time. In light of this analogy and the discussion above regarding the expected failure of the frontier process method for systems with time varying characteristics, one does not expect this method to be useful for analyzing a *time homogeneous* G/G/N+G EDF system off equilibrium. The approach proposed in this paper is to appeal to the MVSM instead.

The aforementioned MSTE and MVSM serve in this work as two main building blocks. The former is used to describe the evolution of the collection of age processes, that keeps track of the time jobs spend in service. The latter captures the EDF discipline, and allows one to express the reneging count process (see Theorem 2.1). The representation of the FME is provided in terms of these two building blocks. The first main difficulty we address is the uniqueness of solutions to the FME (see Section 3.2). Uniqueness cannot be obtained from [13] due to the quite different structure of the FME, as well as to the fact that a certain monotonicity property of the reneging count process with respect to the queue length, used crucially in [13] (see especially (3.14) there), is not clear in our setting. Further, in [2], uniqueness of the FME is established via a minimality result (see Theorem 3.1 there), which holds when the service rate is strictly positive. In a many-server setting the rate at which jobs are transferred from the buffer to the service pool might get arbitrarily close to zero even when the system is busy, and thus the arguments from [2] do not seem to apply. Moreover, unlike the single server EDF model, treated in [2], where the service rate was part of the data, in our setting the rate of job departure from the system (and as a result also the transfer rate alluded to above) is part of the solution, which makes the uniqueness proof significantly more difficult. A second hurdle is the convergence, where several technicalities must be addressed, especially the identification of subsequential limits as solutions of the FME. Our main two results are the uniqueness of FME solutions (Theorem 3.1) and the LLN scale convergence (Theorem 3.2). They are both established under certain assumptions, namely Assumption 3.3 and 3.4, that are treated separately.

As a byproduct of the two main results we obtain a LLN limit for G/G/N+D FCFS (see Corollary 3.2), where 'D' stands for deterministic deadline (this case was not covered by [13,

25, 23]). Let us remark that we could also study scaling limits of other performance measures associated to G/G/N+G EDF, for instance, waiting time. The analysis of waiting time in our settings would be very similar to that available in [15, 13], and therefore we do not pursue it here.

The outline of this paper is as follows. At the end of this section we provide our basic notation. Section 2 describes the model and its scaling, introduces the MVSM and the MSTE and uses them to represent the model dynamics. Section 3 formulates the FME and states the two main results: the uniqueness of solutions of these equations, and the convergence of the scaled processes to their solution. Section 4 provides the proof of uniqueness of solutions. Section 5 establishes tightness of the scaled processes, and finally, Section 6 completes the proof of convergence.

Notation. The set of nonnegative real numbers is denoted by $\mathbb{R}_+$. For $x, y \in \mathbb{R}$, $x \vee y = \max(x, y)$, $x \wedge y = \min(x, y)$, $x^+ = x \vee 0$ and $x^- = (-x) \vee 0$. For $A \subset \mathbb{R}$, $\mathbf{1}_A$ denotes the indicator function of A. Given a metric space $\mathcal{S}$, $\mathbb{C}_b(\mathcal{S})$ and $\mathbb{C}_c(\mathcal{S})$ are, respectively, the space of real-valued bounded continuous functions and the space of real-valued continuous functions with compact support defined on $\mathcal{S}$. When $\mathcal{S}$ is a subset of a finite dimensional vector space, we denote the set of continuously differentiable functions on $\mathcal{S}$ with compact support by $\mathbb{C}^1_c(\mathcal{S})$. Let $\mathbb{C}_{b,+}(\mathcal{S})$ denote the subset of $\mathbb{C}_b(\mathcal{S})$ of $\mathbb{R}_+$-valued functions. Let $\mathbb{D}_{\mathcal{S}}(\mathbb{R}_+)$ and $\mathbb{C}_{\mathcal{S}}(\mathbb{R}_+)$, abbreviated as $\mathbb{D}_{\mathcal{S}}$ and $\mathbb{C}_{\mathcal{S}}$, denote the spaces of functions $\mathbb{R}_+ \to \mathcal{S}$ that are right continuous with finite left limits (RCLL), and respectively, continuous. It is always assumed that $\mathbb{D}_{\mathcal{S}}$ is endowed with the J_1 topology [10]. Denote by $\mathbb{D}^{\uparrow}_{\mathbb{R}_+}$ (resp., $\mathbb{C}^{\uparrow}_{\mathbb{R}_+}$) the subset of $\mathbb{D}_{\mathbb{R}_+}$ (resp., $\mathbb{C}_{\mathbb{R}_+}$) of nondecreasing functions. For $\varphi : \mathbb{R}_+ \to \mathbb{R}$, define $\|\varphi\|_T = \sup_{s\in[0,T]} |\varphi(s)|$ for $T < \infty$, $\|\varphi\|_\infty = \sup_{s\in[0,\infty)} |\varphi(s)|$, and given $\delta > 0$,

$$\mathrm{osc}_\delta(\varphi, T) = \sup\{|\varphi(s) - \varphi(t)| \,:\, |s - t| \le \delta,\ s, t \in [0, T]\}.$$

For functions $\varphi = \varphi(x, t)$ defined on $\mathbb{R}^n \times \mathbb{R}$ we write φ_x and φ_t for the partial derivatives with respect to the first ($x \in \mathbb{R}^n$) and second ($t \in \mathbb{R}$) variable, respectively.

The space of non-negative finite Borel measures on a Polish space $\mathcal{S}$ is denoted by $\mathcal{M}(\mathcal{S})$ and the Borel σ-field of $\mathcal{S}$ is denoted by $\mathcal{B}(\mathcal{S})$. Given $a < \infty$, $\mathcal{M}_a(\mathcal{S})$ denotes the subset of $\mathcal{M}(\mathcal{S})$ consisting of measures with total mass less or equal to a. $\mathcal{M}^0(\mathcal{S})$ denotes the class of atomless measures. We abbreviate $\mathcal{M}(\mathbb{R}_+)$ by $\mathcal{M}$ and $\mathcal{M}^0(\mathbb{R}_+)$ by $\mathcal{M}^0$. For any $\mu \in \mathcal{M}(\mathcal{S})$ and Borel measurable function g on $\mathcal{S}$, denote $\langle g, \mu\rangle = \int g d\mu$. Endow $\mathcal{M}(\mathcal{S})$ with the Prohorov metric, denoted by $d_{\mathcal{M}}$. It is well known that $(\mathcal{M}(\mathcal{S}), d_{\mathcal{M}})$ is a Polish space [8, Appendix], and that this topology on $\mathcal{M}(\mathcal{S})$ is equivalent to the weak topology on this space. Denote by δ_x the unit mass at the point x. Denote convergence in distribution by the symbol '$\Rightarrow$'. Finally, for $\zeta \in \mathbb{D}^{\uparrow}_{\mathbb{R}_+}$, denote by $\mathbf{m}^\zeta$ the Lebesgue-Stieltjes measure that ζ induces on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, namely,

$$\mathbf{m}^\zeta(B) = \zeta(0)\delta_0(B) + \int_{(0,\infty)} \mathbf{1}_B(t) d\zeta_t, \quad B \in \mathcal{B}(\mathbb{R}_+).$$

Throughout, we write "$d\zeta$-a.e." to mean "$d\mathbf{m}^\zeta$-a.e."

2 Model and state dynamics

2.1 Model description

Consider a sequence of systems indexed by $N \in \mathbb{N}$, where the N-th system has N servers that work in parallel. Each system also has a buffer of infinite capacity, in which arriving customers are served according to a non-idling, non-preemptive EDF policy described as follows. Customers arrive with specified deadlines. An arriving customer enters service immediately if there is a free server at the time of its arrival. On the event that all servers are busy, it is queued. When a server becomes available (and the queue is nonempty), it picks the customer that has the earliest deadline among those that are in the queue. Ties are broken according to the order of arrival. Customers that do not enter service by the time of their deadline leave the system (however, customers do not renege while being served). All servers are identical and capable of serving all the customers. The stochastic processes associated with the model, to be introduced below, are all defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The symbol $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$.

A process with right-continuous nondecreasing $\mathbb{Z}_+$-valued sample paths is referred to as a *counting process*. A counting process that has only jumps of size 1 and starts at zero is said to be *simple*. Let E^N be a simple counting process that accounts for arrivals in the N-th system. Namely, the number of arrivals in the time interval $[0, t]$ is given by E^N_t. The jump times of this process, that we denote by $\{a^N_i,\ i \in \mathbb{N}\}$, correspond to the arrival times. This sequence is nondecreasing, so that a^N_i gives the arrival time of the i-th customer. The number of customers in the system (i.e., in service or in the queue) at time 0 is denoted by X^N_0. The sequence $\{a^N_i\}$ is next extended to $\mathcal{I}^N := \{-X^N_0 + 1, -X^N_0 + 2, \ldots, 0\} \cup \mathbb{N}$, so that $\{a^N_i, i \le 0\}$ give the arrival times of the X^N_0 customers present at time $t = 0$. Note that $\mathcal{I}^N$ is a random set of indices.

To model deadlines, let $\{r^N_j, j \in \mathbb{Z}\}$ be an i.i.d. sequence of positive random variables. For $i \in \mathcal{I}^N$, r^N_i represents the *patience time* of the customer i, that is, the deadline of the customer relative to its arrival time. Thus the deadline of customer i is given by $u^N_i = a^N_i + r^N_i$, $i \in \mathcal{I}^N$. We will always refer to u^N_i as the *absolute deadline* of the i-th customer, to avoid confusion with what is referred to in [2] as *relative deadline* (which indicates the deadline with respect to the current time, and is elsewhere referred to as the *remaining patience time* or *lead time* [3, 9, 17]). To recapitulate the reneging rule, customer $i \ge 1$ that does not start service by time u^N_i leaves the system at that time.

Let $\mathcal{Q}^N_0 \in \mathcal{M}$ be a measure describing the state of the queue at time 0, such that for $B \in \mathcal{B}(\mathbb{R}_+)$, $\mathcal{Q}^N_0(B)$ represents the number of those customers in the queue at time 0 whose absolute deadlines are in B. Define a measure-valued process $\mathcal{E}^N$, with sample paths in $\mathbb{D}_{\mathcal{M}}$, as follows. For $B \in \mathcal{B}(\mathbb{R}_+)$ and $t \ge 0$,

$$\mathcal{E}^N_t(B) = \sum_{i=1}^{E^N_t} \delta_{u^N_i}(B) = \sum_{i=1}^{E^N_t} \mathbf{1}_B(u^N_i). \tag{2.1}$$

Then $\mathcal{E}_t(B)$ gives the number of arrivals in $[0, t]$ whose absolute deadlines lie in B. Define α^N, a process with sample paths in $\mathbb{D}_{\mathcal{M}}$, as

$$\alpha^N_t = \mathcal{Q}^N_0 + \mathcal{E}^N_t. \tag{2.2}$$

Introduce the class

$$\mathbb{D}_{\mathcal{M}}^{\uparrow} = \Big\{ \zeta \in \mathbb{D}_{\mathcal{M}} \,:\, \text{the map } t \mapsto \langle f, \zeta_t \rangle \text{ is in } \mathbb{D}_{\mathbb{R}_+}^{\uparrow} \text{ for every } f \in \mathbb{C}_{b,+}(\mathbb{R}_+) \Big\}. \tag{2.3}$$

Since by definition, $t \mapsto \mathcal{E}_t^N(B)$ are nondecreasing, it is easy to see that the sample paths of α^N lie in $\mathbb{D}_{\mathcal{M}}^{\uparrow}$. It is shown in [2, Lemma 2.1] that $\mathbb{D}_{\mathcal{M}}^{\uparrow}$ forms a closed subset of $\mathbb{D}_{\mathcal{M}}$. The MVSM, to be introduced in the next subsection, is defined on this space.

To model service times, consider an i.i.d. sequence, $\{v_i,\, i \in \mathbb{Z}\}$, with common cumulative distribution function G on $[0, \infty)$. For $i \in \mathcal{I}^N$, the service time of customer i is v_i. We assume that G has density, and denote it by g. Let

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

be the right end of the support of g. This constant may be finite or $+\infty$.

We assume that, for each N,

- the arrival process E^N, the sequence of service requirements $\{v_i,\, i \in \mathbb{Z}\}$ and the sequence of patience times $\{r_i^N,\, i \in \mathbb{Z}\}$ are mutually independent.

We refer to the time spent by a customer in service as its *age in service*, or simply its *age* [4, 13, 15]. For a customer $i \in \mathcal{I}^N$ that is ever admitted into service, let γ_i^N denote the admittance time. Let $\gamma_i^N = \infty$ if this customer reneges. Thus, for a customer i initially in service, $\gamma_i^N < 0$. We emphasize that γ_i^N is associated with the i-th customer to arrive into the system rather than the i-th customer to be admitted. The age process ω_i^N associated with customer $i \in \mathcal{I}^N$ is defined as

$$\omega_i^N(t) = (t - \gamma_i^N)^+ \wedge v_i, \qquad t \geq 0.$$

Note that the age of a customer is zero at time t if it has not entered service by that time, and that the age process is identically zero for those customers that renege. Let K^N be a counting process representing cumulative number of admittances into service since time 0 (specifically, $K_0^N = 0$). Since the number of customers initially in the system is given by X_0^N and N is the number of servers, the number of customers initially in service is given by $X_0^N \wedge N$. For $t \in [0, \infty)$, let ν_t^N be the discrete measure on $[0, H^s)$ recording the collection of ages, given by

$$\nu_t^N = \sum_{i=-X_0^N+1}^{E_t^N} \delta_{\omega_i^N(t)} \mathbf{1}_{\{\omega_i^N(t) < v_i,\, t \geq \gamma_i^N\}}. \tag{2.4}$$

Note that $\langle \nu_t^N, 1 \rangle = X_t^N \wedge N$. By construction, the measure-valued process ν_t^N has sample paths in $\mathbb{D}_{\mathcal{M}([0,H^s))}$.

Let D^N be a counting process representing cumulative number of departures from service. We have the explicit representation

$$D_t^N = \sum_{i=-X_0^N+1}^{E_t^N} \sum_{s \in [0,t]} \mathbf{1}_{\{\frac{d\omega_i^N}{dt}(s-) > 0, \frac{d\omega_i^N}{dt}(s+) = 0\}}. \tag{2.5}$$

Let R^N be a counting process representing cumulative number of reneging customers. Let X^N and Q^N be $\mathbb{Z}_+$-valued processes representing the number of customers in the system and in the queue, respectively. Note that the number of customers in service is given by $\langle 1, \nu^N \rangle$, hence $X^N = Q^N + \langle 1, \nu^N \rangle$.

The balance equations for the content of the queue, the service station and the system, respectively, are as follows

$$Q_0^N + E_t^N = Q_t^N + K_t^N + R_t^N, \quad t \geq 0, \tag{2.6}$$

$$X_t^N = Q_t^N + \langle 1, \nu_t^N \rangle, \tag{2.7}$$

$$\langle 1, \nu_0^N \rangle + K_t^N = \langle 1, \nu_t^N \rangle + D_t^N, \quad t \geq 0, \tag{2.8}$$

$$X_0^N + E_t^N = X_t^N + D_t^N + R_t^N, \quad t \geq 0. \tag{2.9}$$

Since the system is working under a nonidling policy, we also have

$$Q_t^N = [X_t^N - N]^+, \quad \text{and } \langle 1, \nu_t^N \rangle = X_t^N \wedge N, \qquad t \geq 0. \tag{2.10}$$

It is also evident from our description that reneging can only occur when all the servers are busy, and so

$$\int_0^{\cdot} [N - X_t^N]^+ dR_t^N = 0. \tag{2.11}$$

We next introduce the filtration that captures the information available as a function of time. Following [13, 15] we introduce a *station process* $s^N = \{s_i^N, i \in \mathcal{I}^N\}$. Assume that the individual servers are labeled by $\{1, \ldots, N\}$. For $t \geq 0$, if customer i has entered service by time t then $s_i^N(t)$ is the number of the server at which it has started (and, possibly, completed) service. Otherwise, $s_i^N(t) = 0$. For $t \in [0, \infty)$, let $\tilde{\mathcal{F}}_t^N$ be the σ-algebra generated by

$$\{X_0^N, \mathcal{Q}_0^N, \mathcal{E}_s^N, \omega_i^N(s), s_i^N(s),\ i \in \mathcal{I}^N,\ s \in [0, t]\},$$

and let $\{\mathcal{F}_t^N\}$ denote the associated right-continuous filtration, complete with respect to $\mathbb{P}$.

2.2 Representation in terms of the MSTE and the MVSM

We introduce the MVSM and then use it to describe the dynamics of the measure-valued process defined above. To this end, we first introduce the *measure-valued Skorohod problem* (MVSP) from [2].

Definition 2.1 (MVSP) *Let $(\boldsymbol{\alpha}, \boldsymbol{\mu}) \in \mathbb{D}_{\mathcal{M}}^{\uparrow} \times \mathbb{D}_{\mathbb{R}_+}^{\uparrow}$. Then $(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\iota}) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}_{\mathcal{M}}^{\uparrow} \times \mathbb{D}_{\mathbb{R}_+}^{\uparrow}$ is said to solve the MVSP for the data $(\boldsymbol{\alpha}, \boldsymbol{\mu})$ if, for each $x \in [0, \infty)$,*

1. $\boldsymbol{\xi}_t[0, x] = \boldsymbol{\alpha}_t[0, x] - \boldsymbol{\mu}_t + \boldsymbol{\beta}_t(x, \infty) + \boldsymbol{\iota}_t$, *for all* $t \geq 0$,
2. $\boldsymbol{\xi}_t[0, x] = 0$, $d\boldsymbol{\beta}_t(x, \infty)$*-a.e.*
3. $\boldsymbol{\xi}_t[0, \infty) = 0$, $d\boldsymbol{\iota}_t$*-a.e.*
4. $\boldsymbol{\beta}_t[0, \infty) + \boldsymbol{\iota}_t = \boldsymbol{\mu}_t$, *for all* $t \geq 0$.

Recall the definition (2.3), and define $\mathbb{C}_{\mathcal{M}}^{\uparrow}$ analogously as a subset of $\mathbb{C}_{\mathcal{M}}$. Let $\mathbb{C}_{\mathcal{M}}^{\uparrow,0}$ denote the collection of members $\zeta \in \mathbb{C}_{\mathcal{M}}^{\uparrow}$ for which the measure ζ_t is atomless for each t. The following was proved in [2, Proposition 2.8 and 2.10].

Proposition 2.1 *1. For every $(\boldsymbol{\alpha}, \boldsymbol{\mu}) \in \mathbb{D}_{\mathcal{M}}^{\uparrow} \times \mathbb{D}_{\mathbb{R}_+}^{\uparrow}$ there exists a unique $(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\iota}) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}_{\mathcal{M}}^{\uparrow} \times \mathbb{D}_{\mathbb{R}_+}^{\uparrow}$ that constitutes a solution to the MVSP with data $(\boldsymbol{\alpha}, \boldsymbol{\mu})$.*

2. Let Θ denote the corresponding solution map, so that $(\boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\iota}) = \Theta(\boldsymbol{\alpha}, \boldsymbol{\mu})$. Then Θ is measurable. Moreover, Θ is continuous on $\mathbb{C}_{\mathcal{M}}^{\uparrow,0} \times \mathbb{C}_{\mathbb{R}_+}^{\uparrow}$.

In addition to Θ defined above we shall use the notation Θ_1, Θ_2 and Θ_3 for denoting the relations $\boldsymbol{\xi} = \Theta_1(\boldsymbol{\alpha}, \boldsymbol{\mu})$, $\boldsymbol{\beta} = \Theta_2(\boldsymbol{\alpha}, \boldsymbol{\mu})$ and $\boldsymbol{\iota} = \Theta_3(\boldsymbol{\alpha}, \boldsymbol{\mu})$.

Remark 2.1 Observe that Definition 2.1 bears close resemblance with the one dimensional Skorohod map. That is, for a given $\psi \in \mathbb{D}_{\mathbb{R}}$, a pair $(\varphi, \eta) \in \mathbb{D}_{\mathbb{R}_+} \times \mathbb{D}_{\mathbb{R}_+}^{\uparrow}$ is said to solve the one-dimensional Skorohod problem for ψ, if $\varphi = \psi + \eta$ and $\varphi(t) = 0$ $d\eta$-a.e. It is well known that there exists a unique pair $(\varphi, \eta) \in \mathbb{D}_{\mathbb{R}+} \times \mathbb{D}_{\mathbb{R}_+}^{\uparrow}$ solving this problem for a given $\psi \in \mathbb{D}_{\mathbb{R}}$. Moreover, if we denote $\Gamma[\psi] = (\varphi, \eta)$, $\Gamma_1[\psi] = \varphi$, $\Gamma_2[\psi] = \eta$, this solution is given by

$$\varphi(t) = \Gamma_1[\psi](t) = \psi(t) - \inf_{s\in[0,t]}(\psi(s) \wedge 0), \quad \text{and} \quad \eta(t) = \Gamma_2[\psi](t) = \varphi(t) - \psi(t).$$

The relation between the two maps Θ and Γ can be made explicit [2, Lemma 2.7], as follows. If $\Theta(\alpha, \mu) = (\xi, \beta, \iota)$ then

$$(\xi[0,x], \beta(x,\infty) + \iota) = \Gamma[\alpha[0,x] - \mu], \quad \text{for all } x \in \mathbb{R}_+. \tag{2.12}$$

Next, for any measurable function φ on $[0, H^s) \times \mathbb{R}_+$ we define a process $D^N(\varphi)$, which takes values in $\mathbb{R}$, by

$$D_t^N(\varphi) = \sum_{i=-X_0^N+1}^{E_t^N} \sum_{s\in[0,t]} \mathbf{1}_{\{\frac{d\omega_i^N}{dt}(s-)>0, \frac{d\omega_i^N}{dt}(s+)=0\}} \varphi(\omega_i^N(s), s). \tag{2.13}$$

From the right continuity of $\{\mathcal{F}_t^N\}$ it follows that $D^N(\varphi)$ is $\{\mathcal{F}^N\}$-adapted. Also, from (2.5) and (2.13) we see that $D^N(1) = D^N$. Three additional measure-valued processes that will be used in our analysis are as follows. Let $\mathcal{Q}^N$ be a process having sample paths in $\mathbb{D}_{\mathcal{M}}$, such that for each $t \in [0,\infty)$, $\mathcal{Q}_t^N(B), B \in \mathcal{B}(\mathbb{R}_+)$, represents the total number of customers in the queue at time t whose absolute deadline is in B. This can be written as

$$\mathcal{Q}_t^N(B) = \sum_{i=-X_0^N+1}^{E_t^N} \mathbf{1}_{\{t<\gamma_i^N \wedge u_i^N\}} \mathbf{1}_B(u_i^N).$$

For $B \in \mathcal{B}(\mathbb{R}_+)$, let

$$\mathcal{K}_t^N(B) = \sum_{i=-X_0^N+1}^{E_t^N} \mathbf{1}_{\{0\le\gamma_i^N\le t\}} \mathbf{1}_B(u_i^N)$$

denote the number of customers admitted into service by time t whose absolute deadline is in B. Note the relation $\mathcal{K}_t^N([0,\infty)) = K_t^N$. It is easy to see that $\mathcal{K}^N$ has sample paths in $\mathbb{D}_{\mathcal{M}}^{\uparrow}$. Similarly, define a process $\mathcal{R}^N$ with sample paths in $\mathbb{D}_{\mathcal{M}}^{\uparrow}$ by

$$\mathcal{R}_t^N(B) = \sum_{i=-X_0^N+1}^{E_t^N} \mathbf{1}_{\{t \geq u_i^N, \gamma_i^N = \infty\}} \mathbf{1}_B(u_i^N).$$

Thus $\mathcal{R}_t^N(B)$ denotes the number of customers that reneged by time t, whose absolute deadlines lie in B. Observe the relation

$$\mathcal{R}_t^N[0,x] = R_{x\wedge t}^N, \quad \text{for all } t,\, x \geq 0.$$

Let

$$\sigma_t^N = \inf\{x \in [0,\infty) \,:\, \mathcal{Q}_t^N[0,x] > 0\}$$

denote the left end of the support of $\mathcal{Q}_t^N$ (defined as $+\infty$ when $\mathcal{Q}_t^N = 0$). Then the fact that customers with deadline within $[0,t]$ cannot be present in the queue at time t is expressed as

$$\mathcal{Q}_t^N[0,t] = 0, \quad \text{for all } t \geq 0, \tag{2.14}$$

and the fact that reneging does not occur prior to the time of the absolute deadline implies

$$\int_0^{\cdot} \mathbf{1}_{\{\sigma_{t-}^N > t\}} dR_t^N = 0. \tag{2.15}$$

Moreover, the prioritization according to absolute deadlines can be expressed by

$$\int_0^{\cdot} \mathcal{Q}_t^N[0,x]\, d\mathcal{K}_t^N(x,\infty) = 0, \qquad x \geq 0, \tag{2.16}$$

and it is also obvious that

$$\int_0^{\cdot} \mathcal{Q}_t^N[0,x]\, d\mathcal{R}_t^N(x,\infty) = 0, \qquad x \geq 0, \tag{2.17}$$

as if there is a customer in the queue with absolute deadline smaller than x at time t, no customer with higher absolute deadline may renege at time t.

The following result identifies crucial relations satisfied by the model, in terms of the MSTE, (2.18) below, and the MVSM, Θ.

Theorem 2.1 *Given $\varphi \in \mathbb{C}_c^1([0,H^s)\times\mathbb{R}_+)$, the processes $\nu^N, D^N(\varphi), K^N, \mathcal{Q}^N, \mathcal{K}^N, \mathcal{R}^N, \alpha^N$ and R^N satisfy*

$$\langle \varphi(\cdot,t), \nu_t^N\rangle = \langle \varphi(\cdot,0), \nu_0^N\rangle + \int_0^t \langle \varphi_x(\cdot,s) + \varphi_s(\cdot,s), \nu_s^N\rangle ds - D_t^N(\varphi) + \int_{[0,t]} \varphi(0,s) dK_s^N, \tag{2.18}$$

and

$$(\mathcal{Q}^N, \mathcal{K}^N + \mathcal{R}^N, 0) = \Theta(\alpha^N, K^N + R^N). \tag{2.19}$$

Proof: Equation (2.18) is shown in [15, Theorem 5.1]. As for (2.19), it follows from the balance equations (2.6)–(2.9) that, for $x \in [0,\infty)$,

$$\mathcal{Q}_t^N[0,x] = \alpha_t^N[0,x] - \mathcal{K}_t^N[0,x] - \mathcal{R}_t^N[0,x] = \alpha_t^N[0,x] - K_t^N - R_t^N + \mathcal{K}^N(x,\infty) + \mathcal{R}^N(x,\infty). \quad (2.20)$$

Defining $\beta^N = \mathcal{K}^N + \mathcal{R}^N$ and using relations (2.16), (2.17) and (2.20) shows that $(\mathcal{Q}^N, \beta^N, 0)$ satisfies all elements of the definition of the MVSP with respect to the data $(\alpha^N, K^N + R^N)$. As a result, (2.19) holds. □

As already mentioned, equation (2.18), that gives the dynamics of the measure-valued process ν^N, originates from [15]. The first term on its RHS is simply an initial condition. The second term accounts for the fact that both age and time increase at rate 1. The remaining two terms correspond to the departure and, respectively, admittance-into-service process.

LLN-scaled version of the processes introduced in this section is attained by downscaling each of them by a factor N. The resulting processes are denoted with a bar, as in $\bar{\mathcal{Q}}^N = N^{-1}\mathcal{Q}^N$, for $N \in \mathbb{N}$. Specifically, $\bar{U}^N = N^{-1}U^N$ for each of the processes $U^N = X^N, Q^N, K^N, R^N, D^N, \mathcal{E}^N, \alpha^N, \nu^N, \mathcal{R}^N, \mathcal{K}^N$.

3 Main results

3.1 Assumptions

We introduce the main assumptions. The first two will be in force throughout this article. Define

$$\mathcal{S}_0 = \Big\{(\mathcal{Q}_0, \nu_0, \alpha, x) \in \mathcal{M}^0 \times \mathcal{M}_1([0,H^s)) \times \mathbb{C}_{\mathcal{M}}^{\uparrow,0} \times \mathbb{R}_+ : \\ 1 - \langle 1, \nu_0\rangle = [1-x]^+,\ \langle 1, \nu_0\rangle + \mathcal{Q}_0(\mathbb{R}_+) = x\Big\}.$$

The assumption regarding convergence of the arrival process is as follows.

Assumption 3.1 *There exists* $(\mathcal{Q}_0, \nu_0, \alpha, X_0) \in \mathcal{S}_0$ *with*

$$\alpha_t[0,x] = \mathcal{Q}_0[0,x] + \int_0^t \mathbf{1}_{\{x \geq s\}} \lambda_s \pi[0, x-s] ds, \quad \forall\, t,\ x \in [0,\infty), \quad (3.1)$$

where $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ *is locally bounded and* π *is a probability measure with* $\pi\{0\} = 0$*, such that, as* $N \to \infty$*,* $\bar{\alpha}^N \Rightarrow \alpha$ *in* $\mathbb{D}_{\mathcal{M}}$*,* $\bar{X}_0^N \Rightarrow X_0$ *in* $\mathbb{R}_+$*, and* $\bar{\nu}_0^N \Rightarrow \nu_0$ *in* $\mathcal{M}_1([0,H^s))$*. Moreover, for any* $T \in (0,\infty)$*,* $\sup_N \mathbb{E}\bar{\alpha}_T^N[0,\infty) < \infty$*.*

The assumption that the limiting measure-valued process α_t takes the form (3.1) corresponds to assuming an asymptotic rate of arrival that follows the function λ_t and an asymptotic patience that is distributed according to the measure π. In particular, the rate of arrivals is allowed to vary with time, but the patience distribution is assumed to be fixed. Working with fixed patience distribution allows us to keep things simple as far as this aspect of the model is concerned. However, extending the results to time-varying patience distribution it is not a serious obstacle, and one could allow that under suitable assumptions. Indeed, this has been done in the single-server setting in [2]; see Assumption 4.5 in [2] for a general structure of patience, and Example 5.3 of [2] for a discussion of this matter.

As for the service time distribution, the hazard rate function of G is defined by

$$h(x) = \frac{g(x)}{1 - G(x)}, \quad x \in [0, H^s).$$

It is easy to see that h is locally integrable on $[0, H^s)$.

Assumption 3.2 *There exists $L^s < H^s$ such that the function h is either bounded or lower-semicontinuous on (L^s, H^s).*

Our results require that either one of the following two assumptions hold.

Assumption 3.3 *1. The density g vanishes at most at finitely many points of $[0, H^s)$. If $\mathcal{Z}$ denotes the set of zeros of g then g is bounded away from zero on any compact subset of $[0, H^s) \setminus \mathcal{Z}$.*

2. $\nu_0 \in \mathcal{M}^0$.

Assumption 3.4 *There exists $\kappa_1 > 0$ such that $\pi[0, \kappa_1] = 0$ where π is given in* (3.1).

Assumption 3.2 was introduced in [15] to guarantee the convergence of sub-sequential limits to a solution of the MSTE. Assumption 3.3 is satisfied by various distributions of practical interest, such as exponential, log-normal, Weibull, gamma, log-logistic etc. When it is in force, we do not make any assumption on the distribution π of patience time. On the other hand, Assmption 3.4 which imposes a condition on π, allows us to treat an extended collection (e.g., uniform over an interval of the form $[a, b]$ where $b > a > 0$, Pareto) of service time distributions. Let us mention that a similar condition to Assumption 3.4 was also used in [17, expression (2.5)] to establish diffusion limits for the G/G/1+G EDF.

3.2 Fluid model equations and statement of results

Recall from Remark 2.1 that $\Gamma = (\Gamma_1, \Gamma_2)$ is the 1-dimensional Skorohod map, that is, for any locally bounded $\psi : \mathbb{R}_+ \to \mathbb{R}$, $\Gamma_1[\psi](t) = \psi(t) - \inf_{s\in[0,t]}[\psi(s) \wedge 0]$, and $\Gamma_2[\psi](t) = -\inf_{s\in[0,t]}[\psi(s) \wedge 0]$, $t \geq 0$.

Definition 3.1 (FME) *A quadruple $(\mathcal{Q}, \nu, K, R)$ in $\mathbb{C}_{\mathcal{M}} \times \mathbb{C}_{\mathcal{M}_1([0,H^s))} \times \mathbb{C}^{\uparrow}_{\mathbb{R}_+} \times \mathbb{C}^{\uparrow}_{\mathbb{R}_+}$ is said to solve the FME with given data $(\mathcal{Q}_0, \nu_0, \alpha, X_0) \in \mathcal{S}_0$, for α as in* (3.1)*, if the initial conditions of $\mathcal{Q}$ and ν are consistent with the data $\mathcal{Q}_0$ and ν_0, one has $\int_0^t \langle h, \nu_s\rangle \, ds < \infty$ for all $t \geq 0$, and, letting*

$$E_t = \alpha_t[0, \infty) - \alpha_0[0, \infty), \tag{3.2}$$

$$Q_t = \mathcal{Q}_t[0, \infty), \tag{3.3}$$

$$X_t = Q_t + \langle 1, \nu_t\rangle, \tag{3.4}$$

and

$$D_t = \int_0^t \langle h, \nu_s\rangle ds, \tag{3.5}$$

the following relations are satisfied. For $\varphi \in \mathbb{C}_c^1([0, H^s) \times [0, \infty))$,

$$\langle \varphi, \nu_t \rangle = \langle \varphi(\cdot, 0), \nu_0 \rangle + \int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \nu_s \rangle ds - \int_0^t \langle h\varphi(\cdot, s), \nu_s \rangle ds + \int_0^t \varphi(0, s) dK_s, \quad (3.6)$$

and

$$Q_0 + E_t = Q_t + K_t + R_t, \quad (3.7)$$

$$Q_t = [X_t - 1]^+, \quad (3.8)$$

$$K_t = \langle 1, \nu_t \rangle - \langle 1, \nu_0 \rangle + D_t \quad (3.9)$$

$$\mathcal{Q} = \Theta_1(\alpha, K + R), \quad (3.10)$$

$$\mathcal{Q}_t[0, t) = 0, \quad \forall\, t \geq 0, \quad (3.11)$$

$$\int_0^{\cdot} [1 - X_s]^+ dR_s = 0, \quad (3.12)$$

$$\sigma_t = t \quad dR\text{-a.e., where for } t \geq 0, \sigma_t = \inf \text{ support}(\mathcal{Q}_t). \quad (3.13)$$

Note that from (3.7), (3.4) and (3.9), it follows that

$$X_t = X_0 + E_t - D_t - R_t. \quad (3.14)$$

The FME are based on analogy to the pre-limit dynamics. Equation (3.6) is similar to (2.18), where the departure term is represented in terms of the hazard rate function. The equations (3.7), (3.4), (3.8) and (3.9) are analogous to (2.6), (2.7), (2.10) and (2.8). Also, (3.11), (3.12) and (3.13) are analogous to (2.14), (2.11) and (2.15), respectively. We see that (3.10) relates the MVSP with data $(\alpha, K + R)$ and therefore analogy comes from (2.19).

Define $\mathcal{E} \in \mathbb{C}_{\mathcal{M}}^{\uparrow,0}$ by

$$\mathcal{E}_t[0, x] = \int_0^t \mathbf{1}_{\{x \geq s\}} \lambda_s \pi[0, x - s] ds, \quad \forall\, t,\ x \in [0, \infty).$$

Then $\alpha = \mathcal{Q}_0 + \mathcal{E}$. Given $t_0 \in [0, \infty)$, define for $s \in [0, \infty)$,

$$\mathcal{E}_s^{[t_0]}[0, x] = \int_0^s \lambda_{t_0+p} \mathbf{1}_{\{x \geq p\}} \pi[0, x - p] dp, \quad \mathcal{Q}_s^{[t_0]}[0, x] = \mathcal{Q}_{t_0+s}[0, t_0 + x], \quad (3.15)$$

$$\nu_s^{[t_0]} = \nu_{t_0+s}, \quad K_s^{[t_0]} = K_{t_0+s} - K_{t_0}, \quad R_s^{[t_0]} = R_{t_0+s} - R_{t_0}.$$

The proof of the following time shift lemma is straightforward.

Lemma 3.1 *Suppose that* $(\mathcal{Q}, \nu, K, R)$ *is a solution to the FME with data* $(\mathcal{Q}_0, \nu_0, \alpha, X_0)$. *Then for any* $t_0 \geq 0$, $(\mathcal{Q}^{[t_0]}, \nu^{[t_0]}, K^{[t_0]}, R^{[t_0]})$ *satisfies the FME with data* $(\mathcal{Q}_{t_0}, \nu_{t_0}, \mathcal{Q}_{t_0} + \mathcal{E}^{[t_0]}, X_{t_0})$.

Let us also recall the following result from [15, Theorem 4.1].

Lemma 3.2 *If* ν *satisfies* (3.6) *then for any* $f \in \mathbb{C}_c(\mathbb{R}_+)$ *we have*

$$\int_{[0,H^s)} f(x)\nu_t(dx) = \int_{[0,H^s)} f(x + t)\frac{1 - G(x + t)}{1 - G(x)}\nu_0(dx) + \int_0^t f(t - s)(1 - G(t - s))dK_s.$$

The first main result of this article states that solutions to the FME are unique.

Theorem 3.1 *Let either Assumption 3.3 or 3.4 hold. Then given* $(\mathcal{Q}_0, \nu_0, \alpha, X_0) \in \mathcal{S}_0$*, there exists at most one solution* $(\mathcal{Q}, \nu, K, R)$ *to the FME.*

Our second main result asserts that the FME characterize the limit.

Theorem 3.2 *Suppose that Assumptions 3.1–3.2 and either Assumption 3.3 or 3.4 hold. Then there exists a unique solution* $(\mathcal{Q}, \nu, K, R)$ *to the FME and the sequence* $(\bar{\mathcal{Q}}^N, \bar{\nu}^N, \bar{K}^N, \bar{R}^N)$ *converges in distribution, as* $N \to \infty$*, to* $(\mathcal{Q}, \nu, K, R)$*.*

We prove Theorem 3.1 in Section 4. The proof is established by two propositions, Proposition 4.1 and 4.2. Sections 5 and 6 are devoted to prove Theorem 3.2, by showing tightness of the scaled processes in Section 5, and the characterization of limits in terms of the FME in Section 6.

In [3, Remark 2.2], the authors establish the fluid limit of $G/M/N + G$ queueing model governed by EDF discipline where the patience time distribution π is assumed to have the property that $x \to \pi[0, x]$ is strictly increasing on the support of π. In view of Theorem 3.1 and Theorem 3.2 we have the following corollary that generalizes the results in [3].

Corollary 3.1 *Let the service requirement distribution be given by an exponential distribution. Assume that Assumptions 3.1 holds and* ν_0 *does not have atoms. Then* $(\bar{\mathcal{Q}}^N, \bar{\nu}^N, \bar{K}^N, \bar{R}^N)$ *converges in distribution, as* $N \to \infty$*, to* $(\mathcal{Q}, \nu, K, R)$ *where* $(\mathcal{Q}, \nu, K, R)$ *is the unique solution to the FME.*

We note that when the patience time distribution is deterministic then G/G/N+D queueing model under EDF policy is same as the queueing model under FCFS (or FIFO) scheduling policy. [13] studies the fluid limit for G/G/N+G queueing systems working under FCFS discipline where it assumes that the patience time distribution has density. This is further generalized in [23] to a more general class of distributions but with continuous patience time distribution. Thus in particular, it does not cover the case of deterministic patience time. Since our Assumption 3.4 includes deterministic patience time we have,

Corollary 3.2 *Let Assumption 3.1-3.2 hold. Then for a queueing model G/G/N+D working under FCFS discipline we have LLN limit and the unique limit can be characterized by the FME* (3.2)-(3.13)*.*

4 Uniqueness of solution to the FME

This section is devoted to the proof of uniqueness of solutions to the FME (3.2)–(3.13), Theorem 3.1. First we show the result under Assumption 3.3 (Proposition 4.1) and then under Assumption 3.4 (in Proposition 4.2).

4.1 Uniqueness under Assumption 3.3

Under Assumption 3.3, for any compact set $\mathbf{C} \subset [0, H^s) \setminus \mathcal{Z}$ we have $\inf_{x\in\mathbf{C}} h(x) > 0$ where h denotes the hazard rate function of G. Now let us begin with the following lemma.

Lemma 4.1 *Let (K, ν) be continuous and satisfy (3.6). Furthermore, assume that the measure ν_0 is atomless. Then ν is atomless, and for any $\varepsilon, T > 0$ there exists $\delta > 0$ such that*

$$\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \nu_t[x-\delta, x+\delta] < \varepsilon.$$

Proof: Fix $s \in [0, T]$ and consider $x \in \mathbb{R}_+$. Since $\nu_0 \in \mathcal{M}^0$, the map $x \mapsto \nu_0((-\infty, x])$ is non-decreasing, bounded and continuous on $\mathbb{R}$, and therefore uniformly continuous on $\mathbb{R}$. Hence there exists $\delta_1 > 0$ such that

$$\sup_{x\in\mathbb{R}_+} \nu_0[x-\delta_1, x+\delta_1] < \frac{\varepsilon}{4}. \tag{4.1}$$

Let $\delta < \delta_1$ and $f_\delta \in \mathbb{C}_{b,+}(\mathbb{R}_+)$ be such that $f_\delta = 1$ on $[x - \frac{\delta}{2}, x + \frac{\delta}{2}]$, f_δ vanishes outside $[x-\delta, x+\delta]$, and $0 \le f_\delta \le 1$. By Lemma 3.2,

$$\langle f_\delta, \nu_s\rangle \le \sup_{x\in\mathbb{R}_+} \nu_0[x-\delta_1, x+\delta_1] + \mathrm{osc}_{2\delta}(K, T).$$

Therefore using the fact that K is continuous and (4.1) we have the result. □

Now we state a uniqueness result under Assumption 3.3.

Proposition 4.1 *Let Assumption 3.3 hold. Let $(\mathcal{Q}^i, \nu^i, K^i, R^i), i = 1, 2$, be two of solutions of the FME (3.2)–(3.13) on $[0, T]$. Then they are equal on $[0, T]$.*

Let $\varepsilon \in (0, 1/2)$. Then by tightness there exists a sequence of compact sets $\mathbf{C}_n$ such that $\max_{i=1,2} \sup_{s\in[0,T]} \nu^i_s(\mathbf{C}^c_n) \to 0$ as $n \to \infty$. Combining with Lemma 4.1 we find a compact set $\mathbf{C} \subset [0, H^s) \setminus \mathcal{Z}$ such that

$$\max_{i=1,2} \sup_{s\in[0,T]} \nu^i_s(\mathbf{C}^c) < \varepsilon. \tag{4.2}$$

Here we have used the fact that $\nu^i([0, H^s)^c) = 0$ which is evident from Lemma 3.2. Define $h_0 = \inf_{x\in\mathbf{C}} h(x)$. From Assumption 3.3(1) we note that h_0 is positive. Let $m = (1-\varepsilon)h_0$ and δ be such that $\sup_{s\in[0,T]} \lambda_s \pi[0, \delta] < m$ and $G(\delta) < \frac{1}{10}$. We shall show that if the two solutions agree at $t < T$ then they agree on the interval $[t, t+\delta]$. This will prove the result. A key observation is that when the solutions are translated using Lemma 3.1, the same compact set $\mathbf{C}$ chosen above works for a fixed ε and thus the choice of δ remains the same on $[0, T]$.

Lemma 4.2 $(\mathcal{Q}^1, \nu^1, K^1, R^1) = (\mathcal{Q}^2, \nu^2, K^2, R^2)$ *on* $[0, \delta]$.

The proof of Lemma 4.2 is based on Lemmas 4.3–4.5 below. Recall D^i from (3.9). Denote $\Delta D = D^1 - D^2$.

Lemma 4.3 *Define $\tilde{R}_s = R^2_s + D^2_s - D^1_s + \langle 1, \nu^2_s\rangle - \langle 1, \nu^1_s\rangle$. Then the image $\tilde{\mathcal{Q}}$ of the data $(\alpha, K^1 + \tilde{R})$ under Θ_1 satisfies $\tilde{\mathcal{Q}} = \mathcal{Q}^2$.*

Proof: By (3.9), $K^1 + \tilde{R} = K^2 + R^2$. As a result, $\tilde{\mathcal{Q}}$ and $\mathcal{Q}^2$ are images of the same data under the map Θ_1, and so they are equal. □

Lemma 4.4 *If for some $t \in [0, \delta)$ and some $z_1 \in (0, \delta - t]$ we have $\mathcal{Q}^1_t[0, t + z_1] = 0$ then $\mathcal{Q}^1_s[0, t + z_1] = 0$ for all $s \in [t, t + z_1)$.*

Proof: Assume the contrary. Then there exist t and z_1 as in the statement of the lemma, and there exists $s \in [t, t+z_1)$ such that $\mathcal{Q}^1_s[0, t+z_1] > 0$. Define $s_0 = \sup\{s' \leq s : \mathcal{Q}^1_{s'}[0, t+z_1] = 0\}$. From (3.10) and [2, Proposition 2.8] we have $\mathcal{Q} \in \mathbb{C}_{\mathcal{M}^0}$. Since $\mathcal{Q}$ is atomless and continuous we have $s_0 \in [t, s)$. Appealing to the 1-dimensional Skorohod map, we have

$$\mathcal{Q}^1_{s'}[0, t + z_1] = \Gamma_1(\mathcal{Q}^1_{s_0}[0, t + z_1] + \bar{\mathcal{E}}[0, t + z_1] - \bar{K}^1 - \bar{R}^1)(s'),$$

where $\bar{\mathcal{E}}_{s'} = \mathcal{E}_{s'} - \mathcal{E}_{s_0}$, $\bar{K}^1_{s'} = K^1_{s'} - K^1_{s_0}$, $\bar{R}^1_{s'} = R^1_{s'} - R^1_{s_0}$. Now by definition we have $\mathcal{Q}_{s'} > 0$ on $(s_0, s]$. Therefore $\langle 1, \nu^1_{s'}\rangle = 1$ on $[s_0, s]$. Hence by (3.9) we have

$$\bar{K}^1_{s'} = \int_{s_0}^{s'} \langle h, \nu^1_p\rangle dp. \tag{4.3}$$

We note that when $\langle 1, \nu^1_p\rangle = 1$, we have $\langle h, \nu^1_p\rangle \geq h_0 \nu^1_p(\mathbf{C}) > h_0(1 - \varepsilon) = m$. Again

$$\bar{\mathcal{E}}_{s'}[0, t + z_1] = \int_{s_0}^{s'} \mathbf{1}_{\{t+z_1 \geq p\}} \lambda_p \pi[0, t + z_1 - p] dp \leq \int_{s_0}^{s'} \sup_{[0,T]} \lambda_s \pi[0, \delta] dp.$$

Therefore $\bar{\mathcal{E}}[0, t + z_1] - \bar{K}^1$ is non-increasing on $[s_0, s]$. Since $\mathcal{Q}^1_{s_0}[0, t + z_1] = 0$ we have $\mathcal{Q}^1_s[0, t + z_1] = 0$ which is a contradiction. Hence follows the lemma. □

Lemma 4.5 *For all $s \in [0, \delta]$ we have $|R^1_s - R^2_s| \leq \|D^1 - D^2\|_s$.*

Proof: Define $\tau = \inf\{t : \ R^1_t > R^2_t + \|D^1 - D^2\|_t\}$. If $\tau \geq \delta$ then there is nothing to prove. Arguing by contradiction, assume that $\tau < \delta$. We claim that there exists $t_1 < \delta$ such that

$$R^1_{t_1} > R^2_{t_1} + \|D^1 - D^2\|_{t_1}, \quad \text{and, for any neighbourhood } \boldsymbol{O} \text{ of } t_1 \text{ we have } \int_{\boldsymbol{O}} dR^1 > 0.$$

To prove this claim we chose $t_2 \in (\tau, \delta)$ such that $R^1_{t_2} > R^2_{t_2} + \|D^1 - D^2\|_{t_2}$. Define $t_1 = \inf\{t \leq t_2, R^1_t = R^1_{t_2}\}$. By (3.2), $E_0 = 0$, and thus by (3.7), $R^1_0 = 0$. Hence we get $t_1 > 0$ as $R^1_{t_1} = R^1_{t_2} > 0$. Again

$$R^1_{t_1} = R^1_{t_2} > R^2_{t_2} + \|D^1 - D^2\|_{t_2} \geq R^2_{t_1} + \|D^1 - D^2\|_{t_1}. \tag{4.4}$$

Also by definition of t_1, we have for any $t < t_1$ that $R^1_{t_1} - R^1_t > 0$. This establishes the claim we made above. Therefore by (3.12) we have a sequence $\{s_n\}, s_n \to t_1$, such that $X^1_{s_n} \geq 1$. Therefore $\langle 1, \nu^1_{s_n}\rangle = 1$ for all n, implying by continuity $\langle 1, \nu^1_{t_1}\rangle = 1$. Define

$$\Delta_0 = R^1_{t_1} - R^2_{t_1} - \|D^1 - D^2\|_{t_1} - \langle 1, \nu^2_{t_1}\rangle + \langle 1, \nu^1_{t_1}\rangle.$$

Then by (4.4), $\Delta_0 > -\langle 1, \nu^2_{t_1}\rangle + \langle 1, \nu^1_{t_1}\rangle$. Since $\langle 1, \nu^2_{t_1}\rangle \le 1$ we have $\Delta_0 > 0$. Since the measures $\mathcal{Q}_0, \mathcal{E}$ are atomless we can find $y = t_1 - \varepsilon_1$, $z = t_1 + \varepsilon_1$ such that $t_1 + \varepsilon_1 < \delta$ and

$$\mathcal{Q}_0(y, z] + \mathcal{E}_{t_1}(y, z] < \frac{\Delta_0}{2}.$$

Define

$$\psi^1_s(z) = \mathcal{Q}_0[0, z] + \mathcal{E}_s[0, z] - K^1_s - R^1_s,$$
$$\tilde{\psi}_s(y) = \mathcal{Q}_0[0, y] + \mathcal{E}_s[0, y] - K^1_s - \tilde{R}_s,$$

where $\tilde{R}$ is same as in Lemma 4.3. From (3.10) we have $\mathcal{Q}^1_t[0, z] = \Gamma_1[\psi^1(z)](t)$ and by (3.11) and Lemma 4.3 we have $0 = \mathcal{Q}^2_{t_1}[0, y] = \Gamma_1[\tilde{\psi}(y)](t_1)$. Now

$$\begin{aligned}\psi^1_{t_1}(z) &= \tilde{\psi}_{t_1}(y) + \mathcal{Q}_0(y, z] + \mathcal{E}_{t_1}(y, z] - (R^1_{t_1} - \tilde{R}_{t_1})\\ &\le \tilde{\psi}_{t_1}(y) + \frac{\Delta_0}{2} - \Big(R^1_{t_1} - R^2_{t_1} - (D^2_{t_1} - D^1_{t_1}) - \langle 1, \nu^2_{t_1}\rangle + \langle 1, \nu^1_{t_1}\rangle\Big).\end{aligned}$$

Noting that $D^2_{t_1} - D^1_{t_1} \le \|D^2 - D^1\|_{t_1}$, it follows from the definition of Δ_0 that

$$\psi^1_{t_1}(z) \le \tilde{\psi}_{t_1}(y) + \frac{\Delta_0}{2} - \Delta_0 = \tilde{\psi}_{t_1}(y) - \frac{\Delta_0}{2}.$$

Since $\Gamma_1[\tilde{\psi}(y)](t_1) = 0$ we have $\tilde{\psi}_{t_1}(y) \le 0$. Let $t_0 = \inf\{t \ge 0 : \psi^1_t(z) \le 0\}$. By the above observation, using the continuity of $t \mapsto \psi^1_t(z)$, we have $t_0 < t_1$. Thus by the definition of t_0, $\psi^1_{t_0}(z) = \inf_{[0,t_0]}(\psi^1_s(z) \wedge 0)$ implying $\mathcal{Q}^1_{t_0}[0, z] = \Gamma_1(\psi^1(z))(t_0) = 0$ and $t_0 < t_1 < z$. Now by Lemma 4.4 we have $\{\sigma^1_t > t\}$ on $[t_0, t_1]$. But dR^1 charges the interval $[t_0, t_1]$. This is a contradiction to (3.13) and thus $\tau \ge \delta$. □

Proof of Lemma 4.2: Denote $\Delta K = K^1 - K^2$. Analogously, define other quantities as $\Delta R, \Delta Q, \Delta X$. Then from (3.7) and (3.8) we have on $[0, \delta]$,

$$\begin{aligned}|\Delta K| \le |\Delta Q| + |\Delta R| &\le |\Delta X| + |\Delta R|\\ &\le 2|\Delta R| + |\Delta D|\\ &\le 3\|D^1 - D^2\|_\delta,\end{aligned}$$

where in the third inequality we used (3.14), and in the last one we used Lemma 4.5. Thus

$$\|\Delta K\|_\delta \le 3\|D^1 - D^2\|_\delta. \tag{4.5}$$

By [15, Corollary 4.4], we have

$$D^i_t = \int_{[0,H^s)} \frac{G(x + t) - G(x)}{1 - G(x)} \nu_0(dx) + \int_0^t g(t - s) K^i_s ds, \quad i = 1, 2.$$

Therefore, for $t \le \delta$,

$$\Delta D_t = \int_0^t g(t - s) \Delta K_s ds,$$

and

$$\begin{aligned}|\Delta D_t| &\le \int_0^t g(t-s)|\Delta K_s|ds \\ &\le \|\Delta K\|_t G(t) \\ &\le \|\Delta K\|_\delta G(\delta).\end{aligned}$$

Hence $\|\Delta D\|_\delta \le \|\Delta K\|_\delta G(\delta) \le \frac{1}{10}\|\Delta K\|_\delta$. Therefore by (4.5), $\|\Delta K\|_\delta \le \frac{3}{10}\|\Delta K\|_\delta$. This shows $\Delta K = 0$ on $[0,\delta]$. Hence $\Delta D = 0$ on $[0,\delta]$. Using Lemma 4.5, also $\Delta R = 0$ on $[0,\delta]$. The equality of $\nu^1 = \nu^2$ on $[0,\delta]$ now follows from Lemma 3.2. Finally, $\mathcal{Q}^1 = \mathcal{Q}^2$ follows from the uniqueness of solutions to the MVSP. □

Proof of Proposition 4.1. To prove uniqueness on $[0,T]$ we argue that if the solutions agree at time $t < T$ then they also agree on $[t,(t+\delta)\wedge T]$ where δ is as in Lemma 4.2. Recall $(\mathcal{Q}^{[t]},\nu^{[t]},K^{[t]},R^{[t]})$ from (3.15). By Lemma 3.1 $(\mathcal{Q}^{[t]},\nu^{[t]},K^{[t]},R^{[t]})$ satisfies the FME with given data $(\mathcal{Q}_t,\nu_t,\mathcal{E}^{[t]},X_t)$ on $[0,T-t]$. Hence uniqueness on $[t,(t+\delta)\wedge T]$ reduces to uniqueness on $[0,(T-t)\wedge\delta]$ with the given data. The latter follows from Lemma 4.2. This completes the proof. □

4.2 Uniqueness under Assumption 3.4

Recall from Assumption 3.4 that $\pi[0,\kappa_1] = 0$. Without loss of generality we assume $\kappa_1 = 2$. We will show that any two solutions agree on $[0,1]$ provided they agree at $t = 0$. Then the uniqueness on any interval follows by applying Lemma 3.1. Let

$$H(x,t) = \mathcal{Q}_0(x,\infty) + \int_0^t \lambda_s \mathbf{1}_{\{x\ge s\}}\pi(x-s,\infty)ds. \tag{4.6}$$

We invoke the Skorohod problem with time-varying boundary. We set the boundary function to be $b_t = H(t,t)$ for $t \ge 0$.

Definition 4.1 *Given $\psi \in \mathbb{C}_{\mathbb{R}_+}$, a pair $(\phi,\eta) \in (\mathbb{C}_{\mathbb{R}_+})^2$, is said to solve the Skorohod problem on the time varying domain $(-\infty,b]$, if*

1. *$\phi_t = \psi_t - \eta_t$, for all $t \ge 0$,*
2. *$\phi_t \le b_t$ for all $t \ge 0$,*
3. *η is non-negative, non-decreasing and $\int_0^\cdot \mathbf{1}_{\{\phi_s<b_s\}}d\eta_s = 0$.*

It is known that there is a unique solution (ϕ,η) given data ψ [5]. It has also been observed in [3] that this type of Skorohod problem shows up in the fluid limits of G/G/1+G queueing models with EDF scheduling. We denote the above Skorohod map by Γ^b, so that $\Gamma_1^b[\psi] = \phi$ and $\Gamma_2^b[\psi] = \eta$. Let $(\mathcal{Q},\nu,K,R)$ be a solution to the FME (3.2)–(3.13). Define

$$\varrho = \inf\{t \ge 0 \ : Q_t \le H(3/2,t)\} \wedge 1.$$

Lemma 4.6 *If $(\mathcal{Q}, \nu, K, R)$ is a solution to the FME (3.2)–(3.13) then*

$$R_s = \Gamma_2^b[Q_0 + E - K](s) \quad \text{for } s \in [0, \varrho), \tag{4.7}$$

$$R_1 - R_\varrho = 0. \tag{4.8}$$

Proof: First we note that for $t \in [0,1]$,

$$b_t = H(t,t) = \mathcal{Q}_0(t,\infty) + \int_0^t \lambda_s \mathbf{1}_{\{t \geq s\}} \pi(t-s,\infty) ds = \mathcal{Q}_0(t,\infty) + E_t,$$

where we used the fact that $\pi[\kappa_1, \infty) = \pi[2,\infty) = 1$. Since $\mathcal{Q}_t$ does not have atoms and $\mathcal{Q}_t[0,t) = 0$ by (3.11) we have for $t \in [0,1]$,

$$\begin{aligned} Q_t = \mathcal{Q}_t(t,\infty) &= Q_t - \mathcal{Q}_t[0,t] \\ &= Q_0 + E_t - K_t - R_t - \Gamma_1[\alpha[0,t] - K - R](t) \\ &= Q_0 + E_t - K_t - R_t - (\alpha[0,t] - K_t - R_t) - \Gamma_2[\alpha[0,t] - K - R](t) \\ &\leq Q_0 + E_t - \mathcal{Q}_0[0,t] - \mathcal{E}_t[0,t] = b_t, \end{aligned} \tag{4.9}$$

where we have used the fact that Γ_2 is non-negative valued and $\mathcal{E}_t[0,t] = 0$. To show (4.7) we may assume $\varrho > 0$, as otherwise there is nothing to prove. Thus on $[0,\varrho)$, $Q_s > H(3/2,s)$ and

$$H(3/2,s) = \mathcal{Q}_0(3/2,\infty) + E_s.$$

For $s \in [0,\varrho)$, we get from (3.7) that

$$0 > H(3/2,s) - Q_s = \mathcal{Q}_0(3/2,\infty) + E_s - \Big(Q_0 + E_s - K_s - R_s\Big) = K_s + R_s - \mathcal{Q}_0[0,3/2]. \tag{4.10}$$

From (4.10) we also note that if for some t we have $K_t + R_t - \mathcal{Q}_0[0,3/2] < 0$ then $K_s + R_s - \mathcal{Q}_0[0,3/2] < 0$ for all $s \leq t$ since $K+R$ is non-decreasing. Therefore we have $\varrho = \sup\{s \in [0,1] : K_s + R_s < \mathcal{Q}_0[0,3/2]\}$. Since $\varrho > 0$ we have $\mathcal{Q}_0[0,3/2] > 0$. Define

$$\tilde{\sigma}_t = \sup\{x \ : \ K_t + R_t \geq \mathcal{Q}_0[0,x]\}.$$

We claim that for $t \in [0,\varrho)$, $\tilde{\sigma}_t$ is equal to the infimum of the support of $\mathcal{Q}_t$, namely, σ_t . Since $\mathcal{Q}_0$ is atomless,

$$K_t + R_t = \mathcal{Q}_0[0,\tilde{\sigma}_t], \quad \text{for } t \in [0,\varrho). \tag{4.11}$$

Fix $t \in [0,\varrho)$. It is enough to show that for any $x < \tilde{\sigma}_t$, we have $\mathcal{Q}_t[0,x] = 0$ and for $x > \tilde{\sigma}_t$, we have $\mathcal{Q}_t[0,x] > 0$. Now $t < \varrho$ implies that $Q_t > H(3/2,t) = \mathcal{Q}_0(3/2,\infty) + E_t$, and therefore by (3.7) we have $\mathcal{Q}_0[0,3/2] > K_t + R_t$. This implies $\tilde{\sigma}_t < 3/2$. Hence it suffices to pick x from $[0,3/2]$. Take $x \in [0,3/2]$ and use (3.10) to obtain

$$\mathcal{Q}_t[0,x] = \mathcal{Q}_0[0,x] - K_t - R_t + \sup_{s \leq t} \Big(K_s + R_s - \mathcal{Q}_0[0,x]\Big)^+. \tag{4.12}$$

Since $K+R$ is non-decreasing, we see from (4.12) that for any $x < \tilde{\sigma}_t$, $\mathcal{Q}_t[0,x] = 0$. Again for $x > \tilde{\sigma}_t$ we have $\sup_{s \leq t} \Big(K_s + R_s - \mathcal{Q}_0[0,x]\Big)^+ = 0$ and therefore $\mathcal{Q}_t[0,x] > 0$. Thus the claim follows. Now for $t \in [0,\varrho)$,

$$b_t - Q_t = H(t,t) - Q_0 - E_t + K_t + R_t = \mathcal{Q}_0(t,\infty) + E_t - Q_0 - E_t + \mathcal{Q}_0[0,\sigma_t]$$

$$= \mathcal{Q}_0(t,\infty) - \mathcal{Q}_0(\sigma_t,\infty),$$

where we use (4.11). Thus $b_t - Q_t > 0$ implies $\sigma_t > t$ for $t \in [0,\varrho)$. Therefore using (3.13) we have for $t \in [0,\varrho)$,

$$\int_0^t \mathbf{1}_{\{Q_s<b_s\}}\, dR_s \le \int_0^t \mathbf{1}_{\{\sigma_s>s\}}\, dR_s = 0. \tag{4.13}$$

Therefore (4.7) follows from (3.7), (4.9) and (4.13). Now we prove (4.8). Without loss of generality, we assume $\varrho < 1$, otherwise there is nothing to prove. Since $\mathcal{Q}_\varrho$ is in $\mathcal{M}^0$ by [2, Proposition 2.8], (4.12) and (4.10) gives us $\mathcal{Q}_\varrho[0,3/2] = \lim_{t\uparrow\rho} \mathcal{Q}_t[0,3/2] = 0$. Thus by definition $\sigma_\varrho = \inf \text{ support}(\mathcal{Q}_\varrho) \ge 3/2$. Therefore from (2.12) and (3.10) we have $K_\varrho + R_\varrho \ge \mathcal{Q}_0[0,3/2]$. Now for any $t \in [\varrho,1]$ and using (4.12) we get

$$\begin{aligned}\mathcal{Q}_t[0,3/2] &= \mathcal{Q}_0[0,3/2] - K_t - R_t + \sup_{s\le t}\big(K_s + R_s - \mathcal{Q}_0[0,3/2]\big)^+ \\ &= \mathcal{Q}_0[0,3/2] - K_t - R_t + \big(K_t + R_t - \mathcal{Q}_0[0,3/2]\big) = 0.\end{aligned}$$

Thus for $t \in [\varrho,1]$ we have $\sigma_t \ge 3/2 > t$. Hence from (3.13) we obtain $R_1 - R_\varrho = 0$. □

Proposition 4.2 *Let Assumption 3.4 hold. If $(\mathcal{Q}^i,\nu^i,K^i,R^i), i = 1,2$, solve the FME (3.2)–(3.13) on $[0,T]$ with data $(\mathcal{Q}_0,\nu_0,\alpha,X_0)$ then $(\mathcal{Q}^1,\nu^1,K^1,R^1) = (\mathcal{Q}^2,\nu^2,K^2,R^2)$ on $[0,T]$.*

Proof: As we commented earlier, we show the uniqueness on $[0,1]$ (keep in mind that $\kappa_1 = 2$) and then one can apply Lemma 3.1 to extend the time interval. We distinguish between two cases.

Case 1. $\mathcal{Q}_0[0,3/2] > 0$. Then $\varrho > 0$ where ϱ is as in Lemma 4.6. Let ϱ^i be the time corresponding to the i-th system. As shown in Lemma 4.6, we have that for $s \in [0,\varrho^i)$, $K^i_s + R^i_s < \mathcal{Q}_0[0,3/2]$. Thus for $i = 1,2$, on $[0,\varrho^i)$, we have

$$Q^i_s = Q_0 + E_s - K^i_s - R^i_s > \mathcal{Q}_0(3/2,\infty) + E_s \ge 0. \tag{4.14}$$

Thus the system has a busy period in $[0,\varrho^i]$. We apply [15, Corollary 4.4] by which

$$K^i_s = \langle 1,\nu^i_s\rangle - \langle 1,\nu_0\rangle + \int_{[0,H^s)} \frac{G(x+s)-G(x)}{1-G(x)}\nu_0(dx) + \int_0^t g(t-s)K^i_s ds. \tag{4.15}$$

On $[0,\varrho^i]$ we have $\langle 1,\nu^i_s\rangle = 1$ by (4.14) and therefore from (4.15) we obtain

$$K^i_s = \int_{[0,H^s)} \frac{G(x+s)-G(x)}{1-G(x)}\nu_0(dx) + \int_0^t g(t-s)K^i_s ds, \quad \text{for } s \in [0,\varrho^i].$$

However, the above is a renewal equation and therefore has a unique solution [1, Th 5.2.4] Therefore $K^1 = K^2$ on $[0,\varrho^1 \wedge \varrho^2]$. Thus by Lemma 3.2 we have $\nu^1 = \nu^2$ on $[0,\varrho^1 \wedge \varrho^2]$. Applying Lemma 4.6 we obtain $(\mathcal{Q}^1,\nu^1,K^1,R^1) = (\mathcal{Q}^2,\nu^2,K^2,R^2)$ on $[0,\varrho^1 \wedge \varrho^2]$. Again, this would contradict the definition of ϱ unless we have $\varrho^1 = \varrho^2$. It remains to prove that the equality holds on $[\varrho,1]$. We already know from Lemma 4.6 that $R^i_1 - R^i_\varrho = 0$ for $i = 1,2$. Thus to show equality we only need to show that $(\nu^1,K^1) = (\nu^2,K^2)$ on $[\varrho,1]$. To this end, we consider the shifted solutions $(\nu^{i,[\varrho]},K^{i,[\varrho]})$ for $i = 1,2$, defined as in (3.15). Since on $[\varrho,1]$ the system behaves like a system without reneging the equality follows from [15, Theorem 4.6].

Case 2. Let $\mathcal{Q}_0[0,3/2] = 0$. In this case $\varrho^i = 0$, and by Lemma 4.6, $R^i_1 = 0$ for $i = 1,2$. The equality now follows by similar arguments to those of case 1. □

5 Tightness of the scaled processes

In this section we establish tightness of the scaled processes defined in Section 3. First we introduce a family of martingales that plays a crucial role in this proof (following ideas of [15, 13]). For any bounded measurable function φ defined on $[0, H^s) \times \mathbb{R}_+$, let

$$A_t^N(\varphi) = \int_0^t \int_{[0,H^s)} \varphi(x,s) h(x) \nu_s^N(dx)\, ds. \tag{5.1}$$

Lemma 5.1 *For every bounded measurable function φ defined on $[0, H^s) \times \mathbb{R}_+$, the process $\mathcal{M}^N(\varphi)$ defined by*

$$\mathcal{M}^N(\varphi) = D^N(\varphi) - A^N(\varphi), \tag{5.2}$$

where $D^N(\varphi)$ is given by (2.13), is a local $\mathcal{F}_t^N$-martingale. Moreover, for every $N \in \mathbb{N}$, $t \in [0,\infty)$ and $c \in [0, H^s)$,

$$\left|A_t^N(\varphi)\right| \le \|\varphi\|_\infty (X_0^N + E_t^N) \int_0^c h(x)dx < \infty, \tag{5.3}$$

for any $\varphi \in \mathbb{C}_c([0, H^s) \times \mathbb{R}_+)$ with support$(\varphi) \subset [0,c] \times \mathbb{R}_+$. In addition, the predictable quadratic variation process $\langle \bar{\mathcal{M}}^N(\varphi) \rangle$ of $\bar{\mathcal{M}}^N(\varphi) = \frac{1}{N}\mathcal{M}^N(\varphi)$ satisfies

$$\lim_{N\to\infty} \sup_{s\in[0,t]} \mathbb{E}[\langle \bar{\mathcal{M}}_s^N(\varphi) \rangle] = 0, \quad \bar{\mathcal{M}}^N(\varphi) \Rightarrow 0, \tag{5.4}$$

for every $\varphi \in \mathbb{C}_b([0, H^s) \times \mathbb{R}_+)$.

Proof: The proof of the local martingale property of $\mathcal{M}^N(\varphi)$ follows using the argument in [15, Lemma 5.4 and Corollary 5.5]. The filtration used in [15] is smaller than the one considered here. We have one added element in our filtration $\{\mathcal{F}_t^N\}$, corresponding to the deadline information. But the deadlines u^N are independent of the service requirement process. Therefore one can apply conditional independence in a straightforward fashion to obtain the local martingale property of $\mathcal{M}^N(\varphi)$. A similar argument by independence is also used in [13, Proposition 5.1], where one can check for the calculations. The proof of (5.3) can be established following [15, Proposition 5.7]. For the proof of (5.4) we refer to [15, Lemma 5.9]. □

A sequence of processes with sample paths in $\mathbb{D}_{\mathcal{S}}$, $\mathcal{S}$ being a Polish space, is said to be $\mathbb{C}$-*tight* if it is tight in $\mathbb{D}_{\mathcal{S}}$ (in the J_1 topology) and, in addition, any subsequential limit has, a.s., paths in $\mathbb{C}_{\mathcal{S}}$. The following characterization will be useful.

Characterization of $\mathbb{C}$-tightness for processes with sample paths in $\mathbb{D}_{\mathbb{R}}$ (see Proposition VI.3.26 of [12]): *$\mathbb{C}$-tightness of a sequence of processes Z^N is equivalent to*

C1. The sequence of random variables $\|Z^N\|_T$ is tight for every fixed $T < \infty$, and

C2. For every $T < \infty$, $\varepsilon > 0$ and $\eta > 0$ there exist N_0 and $\theta > 0$ such that

$$N \ge N_0 \text{ implies } \mathbb{P}(osc_\theta(Z^N, T) > \eta) < \varepsilon.$$

Lemma 5.2 *Let Assumption 3.1 hold. Then, for $\bar{Z}^N = \bar{K}^N, \bar{X}^N, \bar{R}^N, \langle 1, \bar{\nu}^N \rangle$, the sequence $\{\bar{Z}^N\}$ and the sequences $\{\bar{D}^N(\varphi)\}, \{\bar{A}^N(\varphi)\}$, for every $\varphi \in \mathbb{C}_{b,+}([0, H^s) \times \mathbb{R}_+)$, are $\mathbb{C}$-tight.*

Proof: Fix $T \in (0,\infty)$. By Lemma 5.1 we have for some increasing sequence of stopping times $\hat{\tau}_n$, $\hat{\tau}_n \to \infty$, that $\mathbb{E}[\bar{A}^N_{T\wedge\hat{\tau}_n}(\varphi)] = \mathbb{E}[\bar{D}^N_{T\wedge\hat{\tau}_n}(\varphi)]$. Therefore by monotone convergence, $\mathbb{E}[\bar{A}^N_T(\varphi)] = \mathbb{E}[\bar{D}^N_T(\varphi)]$. Using the final assertion of Assumption 3.1,

$$\sup_N \mathbb{E}[\bar{A}^N_T(\varphi)] = \sup_N \mathbb{E}[\bar{D}^N_T(\varphi)] \le \|\varphi\|_\infty \sup_N \mathbb{E}[\bar{X}^N_0 + \bar{E}^N_T] < \infty. \tag{5.5}$$

Similarly, using (2.6)–(2.10) we obtain

$$\sup_N \mathbb{E}[\bar{U}^N_T] \le \sup_N \mathbb{E}[\bar{X}^N_0 + \bar{E}^N_T] < \infty, \tag{5.6}$$
$$\text{for } \bar{U}^N_T = \bar{K}^N_T, \bar{R}^N_T, \sup_{[0,T]}\langle 1, \bar{\nu}^N\rangle, \sup_{[0,T]} \bar{X}^N.$$

Since $\bar{A}^N_T(\varphi), \bar{D}^N_T(\varphi), \bar{K}^N, \bar{R}^N$ are non-decreasing for $\varphi \in \mathbb{C}_{b,+}([0,H^s)\times\mathbb{R}_+)$, criterion $C1$ follows from (5.5) and (5.6). Now we show that criterion C2 also holds. We start with $\bar{R}^N$. Fix $\varepsilon, \eta \in (0,\infty)$. Let $\theta > 0$ and $s,t \in [0,T], s < t$ be such that $t \le s + \theta$. We estimate $R^N_t - R^N_s$. Recall $\mathcal{E}^N$ from (2.1). Then

$$\begin{aligned} R^N_t - R^N_s &\le \mathcal{Q}^N_s[s,t] + \mathcal{E}^N_t[s,t] - \mathcal{E}^N_s[s,t] \\ &\le \alpha^N_s[s,t] + \mathcal{E}^N_t[t-\theta,t]. \end{aligned}$$

By Assumption 3.1 we have $\sup_{t\in[0,T]} d_{\mathcal{M}}(\bar{\alpha}^N_t, \alpha_t) \Rightarrow 0$ and $\sup_{t\in[0,T]} d_{\mathcal{M}}(\bar{\mathcal{E}}^N_t, \mathcal{E}_t) \Rightarrow 0$ as $N \to \infty$. Since $\alpha \in \mathbb{C}^{\uparrow,0}_{\mathcal{M}}$, we see that $x \mapsto \alpha_t[0,x]$ is uniformly continuous on $\mathbb{R}_+$, uniformly with respect to $t \le T$. On the other hand we also have

$$d_{\mathcal{M}}(\bar{\alpha}^N_t, \alpha_t) \le \sup_{x\in[0,\infty)} |\bar{\alpha}^N_t[0,x] - \alpha_t[0,x]| \le d_{\mathcal{M}}(\bar{\alpha}^N_t, \alpha_t) + \mathrm{osc}_{2d_{\mathcal{M}}(\bar{\alpha}^N_t,\alpha_t)}(\alpha_t[0,\cdot], T).$$

Similar fact also holds true for $\bar{\mathcal{E}}^N, \mathcal{E}$. Therefore we can find $\theta > 0$ so that

$$\lim_{N\to\infty} \mathbb{P}\Big(\sup_{s\in[0,T]} \bar{\alpha}^N_s[s,s+\theta] + \sup_{t\in[0,T]} \bar{\mathcal{E}}^N_t[t-\theta,t] > \eta\Big) = 0.$$

Combining the above two displays we see that $\bar{R}^N$ satisfies criterion C2. Now, from [15, Lemma 5.8(2), Lemma 5.12] we obtain that, for any $\varphi \in \mathbb{C}_{b,+}([0,H^s)\times\mathbb{R}_+)$,

$$\lim_{\delta\to 0} \limsup_{N\to\infty} \mathbb{E}\Big[\sup_{t\in[0,T]} (\bar{A}^N_{t+\delta}(\varphi) - \bar{A}^N_t(\varphi))\Big] = 0. \tag{5.7}$$

Combining (5.7) with (5.4) we see that both the sequences $\{\bar{D}^N(\varphi)\}, \{\bar{A}^N(\varphi)\}$ satisfy C2 for every $\varphi \in \mathbb{C}_{b,+}(\mathbb{R}_+)$. In particular, $\{\bar{D}^N(1)\} = \{\bar{D}^N\}$ also satisfies C2. Since

$$\begin{aligned} |\bar{X}^N_t - \bar{X}^N_s| &\le |\bar{E}^N_t - \bar{E}^N_s| + |\bar{D}^N_t - \bar{D}^N_s| + |\bar{R}^N_t - \bar{R}^N_s|, \\ |\langle 1, \bar{\nu}^N_t\rangle - \langle 1, \bar{\nu}^N_s\rangle| &\le |\bar{X}^N_t - \bar{X}^N_s|, \end{aligned}$$

using (2.9) and (2.10), we see that $\{\bar{X}^N\}$ and $\{\langle 1, \bar{\nu}^N\rangle\}$ also satisfy C2. finally, the sequence $\bar{K}^N$ satisfies C2 by (2.8). □

Recall the metric $d_{\mathcal{M}}$ on $\mathcal{M}$ from our notation.

Lemma 5.3 *Suppose that Assumption 3.1 holds. For every ε, η and $T \in (0,\infty)$ there exist $\delta \in (0,\infty)$ and N_0 such that*

$$\text{for all } N \geq N_0, \quad \mathbb{P}\Big(\sup_{0\leq s\leq t\leq s+\delta\leq T} d_{\mathcal{M}}(\bar{\nu}_s^N, \bar{\nu}_t^N) > \eta \Big) < \varepsilon.$$

Proof: Recall the definition (2.4) of ν_t^N. Let $F \subset \mathbb{R}_+$ be any closed set and denote by F^{ε_1} its ε_1-enlargement in $\mathbb{R}_+$, for a given $\varepsilon_1 \in (0,\eta)$. Then for $s \leq t \leq s + \varepsilon_1/2$ we have from (2.4)

$$\begin{aligned}
&\bar{\nu}_t^N(F) - \bar{\nu}_s^N(F^{\varepsilon_1}) \qquad (5.8)\\
&= \frac{1}{N} \sum_{j=-X_0^N+1}^{E_t^N} \mathbf{1}_F(\omega_j^N(t)) \mathbf{1}_{\{\omega_j^N(t)<v_j,\ t\geq\gamma_i^N\}} - \frac{1}{N} \sum_{j=-X_0^N+1}^{E_s^N} \mathbf{1}_{F^{\varepsilon_1}}(\omega_j^N(s)) \mathbf{1}_{\{\omega_j^N(s)<v_j,\ s\geq\gamma_i^N\}}\\
&\leq \frac{1}{N} \sum_{j=-X_0^N+1}^{E_s^N} \Big(\mathbf{1}_F(\omega_j^N(t)) \mathbf{1}_{\{\omega_j^N(t)<v_j,\ s\geq\gamma_i^N\}} - \mathbf{1}_{F^{\varepsilon_1}}(\omega_j^N(s)) \mathbf{1}_{\{\omega_j^N(s)<v_j,\ s\geq\gamma_i^N\}} \Big) + \bar{K}_t^N - \bar{K}_s^N\\
&\leq \bar{K}_t^N - \bar{K}_s^N,
\end{aligned}$$

where the last inequality follows from the following fact: if a customer j has entered service by time s and receives service at time t with $\omega_j^N(t) \in F$, then by definition $\omega_j^N(s) \in F^{\varepsilon_1}$ as $t - s \leq \varepsilon_1/2$ and ω_j^N grows linearly. Also

$$\begin{aligned}
&\bar{\nu}_s(F) - \bar{\nu}_t(F^{\varepsilon_1}) \qquad (5.9)\\
&\leq \frac{1}{N} \sum_{j=-X_0^N+1}^{E_s^N} \mathbf{1}_F(\omega_j^N(s)) \mathbf{1}_{\{\omega_j^N(s)<v_j,\ s\geq\gamma_i^N\}} - \frac{1}{N} \sum_{j=-X_0^N+1}^{E_s^N} \mathbf{1}_{F^{\varepsilon_1}}(\omega_j^N(t)) \mathbf{1}_{\{\omega_j^N(t)<v_j,\ s\geq\gamma_i^N\}}\\
&= \frac{1}{N} \sum_{j=-X_0^N+1}^{E_s^N} \Big(\mathbf{1}_F(\omega_j^N(s)) \mathbf{1}_{\{\omega_j^N(s)<v_j,\ s\geq\gamma_i^N\}} - \mathbf{1}_{F^{\varepsilon_1}}(\omega_j^N(t)) \mathbf{1}_{\{\omega_j^N(t)<v_j,\ s\geq\gamma_i^N\}} \Big)\\
&\leq \bar{D}_t^N - \bar{D}_s^N,
\end{aligned}$$

where the last inequality follows from the fact that if a customer is in service at time s but not at time t then it must have completed its service in the time interval $(s,t]$. Now using Lemma 5.2 we have $\theta \in (0,\infty), N_0 \in \mathbb{N}$ such that

$$\text{for all } N \geq N_0, \quad \mathbb{P}(\mathrm{osc}_\theta(\bar{Z}, T) > \varepsilon_1) < \varepsilon/2, \quad \text{for } \bar{Z} = \bar{K}^N, \bar{D}^N. \qquad (5.10)$$

Since osc_θ is increasing in θ we can chose $\theta \in (0, \varepsilon_1/2)$. Thus combining (5.8), (5.9) and (5.10) we obtain

$$\text{for all } N \geq N_0, \quad \mathbb{P}\Big(\sup_{0\leq s\leq t\leq t+\delta\leq T} d_{\mathcal{M}}(\bar{\nu}_s^N, \bar{\nu}_t^N) > \eta \Big) < \varepsilon,$$

for $\delta = \theta$. This completes the proof. □

Lemma 5.4 *Suppose that Assumption 3.1 holds. Then the sequence $\{\bar{\nu}^N\}$ is C-tight.*

Proof: First we argue that tightness holds for $\{\bar{\nu}^N\}$, then $\mathbb{C}$-tightness. As far as tightness is concerned, since we have already established the oscillation bound stated in Lemma 5.3, to show tightness we only need to show the compact containment property (see [10, Corollary 3.7.4]), i.e., that for each $T, \eta > 0$, there exists a compact set $\mathbf{K}_{T,\eta} \subset \mathcal{M}$ such that

$$\liminf_{N\to\infty} \mathbb{P}\left(\bar{\nu}_t^N \in \mathbf{K}_{T,\eta} \text{ for all } t \in [0,T]\right) > 1 - \eta.$$

The proof of this statement follows just as in [15, Lemma 5.12].

Next we show $\mathbb{C}$-tightness. Define for $\zeta \in \mathbb{D}_{\mathcal{M}}$,

$$J(\zeta) = \int_0^\infty e^{-t}[J(\zeta,t) \wedge 1]dt, \quad \text{where } J(\zeta,t) = \sup_{s\le t} d_{\mathcal{M}}(\zeta_s, \zeta_{s-}).$$

By [10, Theorem 3.10.2], to show $\mathbb{C}$-tightness it suffices to show that for any $\varepsilon, \eta > 0$

$$\liminf_{N\to\infty} \mathbb{P}(J(\bar{\nu}^N) \le \varepsilon) \ge 1 - \eta.$$

However, this is obvious from Lemma 5.3. This completes the proof. □

Lemma 5.5 *Let Assumption 3.1 hold. Then the collection of measure-valued processes* $\{\bar{\mathcal{Q}}^N\}$ *is* $\mathbb{C}$*-tight.*

Proof: Note by (2.19) that $\bar{\mathcal{Q}}^N$ is the image of $\bar{\alpha}^N$ and $\bar{K}^N + \bar{R}^N$ under Θ. Since we have already proved $\mathbb{C}$-tightness of $\bar{K}^N$ and $\bar{R}^N$, the result is an immediate consequence of the continuity of Θ on $\mathbb{C}^{\uparrow,0}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$, shown in [2, Proposition 2.10]. □

Next, introduce two measure-valued processes associated to D^N and its compensator, taking values in $\mathcal{M}([0,H^s) \times R_+)$. For A a measurable subset of $[0,H^s) \times R_+$, let

$$\bar{\mathcal{D}}_t^N(A) = \bar{D}_t^N(\mathbf{1}_A),$$
$$\bar{\mathcal{A}}_t^N(A) = \bar{A}_t^N(\mathbf{1}_A),$$

where $D_t^N(\varphi)$ and $A_t^N(\varphi)$ are given by (2.13) and (5.1), respectively. For $\varphi \in \mathbb{C}_c([0,H^s) \times R_+)$, denote $\bar{\mathcal{D}}_t^N(\varphi) = \bar{\mathcal{D}}_t^N(\varphi)$ and $\bar{\mathcal{A}}_t^N(\varphi) = \bar{\mathcal{A}}_t^N(\varphi)$. Writing $\varphi = \varphi^+ - \varphi^-$ and using Lemma 5.2 we have the sequences $\{\bar{D}^N(\varphi)\}, \{\bar{A}^N(\varphi)\}$ $\mathbb{C}$-tight, for every $\varphi \in \mathbb{C}_b([0,H^s) \times \mathbb{R}_+)$. Using (5.5) and the arguments of [15, Lemma 5.13] one can show that the processes $\bar{\mathcal{D}}^N, \bar{\mathcal{A}}^N$ satisfy the compact containment condition in the sense of Jakubowski. Thus we can establish Jakubowski's criteria for compactness of processes in $\mathbb{D}_{\mathcal{M}}([0,H^s) \times \mathbb{R}_+)$ for $\bar{\mathcal{D}}^N, \bar{\mathcal{A}}^N$ (see [15, Lemma 5.13]) and obtain the following result.

Lemma 5.6 *Let Assumption 3.1 hold. Then the sequences* $\{\bar{\mathcal{D}}^N\}$ *and* $\{\bar{\mathcal{A}}^N\}$ *are tight in the space* $\mathbb{D}_{\mathcal{M}([0,H^s)\times\mathbb{R}_+)}$.

6 Characterization of limits

Finally, we prove Theorem 3.2. This section is devoted in the characterization of the subsequential limits of the scaled processes. Given the tightness result from section 5 and the uniqueness of solutions to the FME, it suffices to prove that any subsequential limit of the scaled processes solves the FME. In other words, Theorem 3.2 is an immediate consequence of the following.

Proposition 6.1 *Let the hypotheses of Theorem 3.2 hold. If $(\mathcal{Q}, \nu, K, R)$ is any subsequential limit of $(\bar{\mathcal{Q}}^N, \bar{\nu}^N, \bar{K}^N, \bar{R}^N)$ then it solves the FME (3.2)–(3.13) with data $(\mathcal{Q}_0, \nu_0, \alpha, X_0)$.*

We define

$$\mathcal{Y} = \mathbb{R}_+ \times (\mathbb{D}_{\mathbb{R}_+})^3 \times (\mathbb{D}_{\mathcal{M}})^2 \times \mathcal{M}_1([0, H^s)) \times \mathbb{D}_{\mathcal{M}_1([0,H^s))} \times (\mathbb{D}_{\mathcal{M}([0,H^s)\times\mathbb{R}_+)})^2.$$

We equip $\mathcal{Y}$ with the product topology. Let

$$\bar{Y}^N = (\bar{X}_0^N, \bar{X}^N, \bar{K}^N, \bar{R}^N, \bar{\alpha}^N, \bar{\mathcal{Q}}^N, \bar{\nu}_0, \bar{\nu}^N, \bar{\mathcal{A}}^N, \bar{\mathcal{D}}^N) \in \mathcal{Y}.$$

Applying Lemma 5.1 –5.6 we see that $\bar{Y}^N$ is a tight sequences and thus it has a convergent subsequence. Let Y be one of the subsequential limit of $\bar{Y}^N$. In fact, Y would be of following form because of Lemma 5.1,

$$Y = (X_0, X, K, R, \alpha, \mathcal{Q}, \nu_0, \nu, \mathcal{A}, \mathcal{A}), \quad \text{where } \mathcal{A} \in \mathbb{D}_{\mathcal{M}([0,H^s)\times\mathbb{R}_+)}.$$

Also because of our $\mathbb{C}$-tightness we have $X, K, R, \mathcal{Q}, \nu$ continuous. Moving to the subsequence and applying Skorohod representation theorem we can assume that $\bar{Y}^N \to Y$ a.s. as $N \to \infty$ on some probability space, say $(\Omega, \mathcal{F}, \mathbb{P})$.

The following result follows from [15, Proposition 5.17]

Lemma 6.1 *Suppose Assumption 3.1 and 3.2 holds. Then for every $\varphi \in \mathbb{C}_b([0, H^s) \times \mathbb{R}_+)$,*

$$\mathcal{A}_t(\varphi) = \int_0^t \langle \varphi(\cdot, s) h(\cdot), \nu_s \rangle \, ds, \quad t \in [0, \infty).$$

In view of Lemma 6.1 following relations hold a.s.: for any $\varphi \in \mathbb{C}_c^1([0, H^s) \times [0, \infty))$,

$$\langle \varphi, \nu_t \rangle = \langle \varphi(\cdot, 0), \nu_0 \rangle + \int_0^t \langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \nu_s \rangle ds - \int_0^t \langle h\varphi(\cdot, s), \nu_s \rangle ds + \int_0^t \varphi(0, s) dK_s, \quad (6.1)$$

where with $Q_t = \mathcal{Q}_t(\mathbb{R}_+)$, we have

$$Q_0 + E_t = Q_t + K_t + R_t, \quad (6.2)$$

$$X_t = Q_t + \langle 1, \nu_t \rangle, \quad (6.3)$$

$$Q_t = [X_t - 1]^+, \quad (6.4)$$

$$K_t = \langle 1, \nu_t \rangle - \langle 1, \nu_0 \rangle + D_t = \langle 1, \nu_t \rangle - \langle 1, \nu_0 \rangle + \int_0^t \langle h, \nu_s \rangle ds, \quad (6.5)$$

where (6.1) follows from (2.18), (6.2)-(6.5) follows from (2.6)-(2.8), and (6.4) follows from (2.10). From (2.14) and (2.11) we get

$$\mathcal{Q}_t[0, t) = 0, \quad (6.6)$$

$$\int_0^\cdot [1 - X_s]^+ dR_s = 0. \quad (6.7)$$

Since $\mathcal{Q}_t^N[a,b] \le \alpha_t^N[a,b]$ we obtain from Assumption 3.1 that $\mathcal{Q}_t$ does not have any atoms for every $t \in [0,\infty)$ a.s. Therefore from Proposition 2.1 and Theorem 2.1 we get that, a.s., for all $x \in [0,\infty)$,

$$\mathcal{Q}_t[0,x] = \Gamma_1[\alpha[0,x] - K - R](t), \quad t \ge 0. \tag{6.8}$$

Thus to complete the proof of Proposition 6.1, it remains to show (3.13). Define

$$\sigma_t = \inf \text{ support}(\mathcal{Q}_t) = \inf\{x \in \mathbb{R}_+ \ : \ \mathcal{Q}_t[0,x] > 0\}.$$

From (2.14) we have $\mathcal{Q}_t[0,t) = 0$ and therefore $\sigma_t \ge t$ for all $t \in [0,\infty)$. We have the following result which is in analogy with (2.15).

Lemma 6.2 *Let either Assumption 3.3 or Assumption 3.4 hold. Then we have a.s.,*

$$\int_0^{\cdot} \mathbf{1}_{\{\sigma_s > s\}} dR_s = 0.$$

Proof: Fix $T > 0$. We show that

$$\int_0^{T} \mathbf{1}_{\{\sigma_s > s\}} dR_s = 0.$$

Since $\{\sigma_s > s\} = \cup_{n\in\mathbb{N}}\{\sigma_s > s + \frac{1}{n}\}$, it is enough to show that for any positive $\hat{\delta}$ we have a.s.,

$$\int_0^{T} \mathbf{1}_{\{\sigma_s > s+\hat{\delta}\}} dR_s = 0. \tag{6.9}$$

Notice that at this stage of the proof it has not yet been established that the subsequential limit forms a solution to the FME, and therefore we cannot treat it as deterministic.

The measurability of the set

$$A_0 = \left\{ \int_0^{T} \mathbf{1}_{\{\sigma_s > s+\hat{\delta}\}} dR_s = 0 \right\}$$

can be shown following the arguments in [2, Lemma 5.9]. That is, from (6.8) one can easily show that $t \mapsto \mathcal{Q}_t[0,t+a]$ is right continuous for every $a \ge 0$ (see also [2, Lemma 4.4]). Therefore the map $(\omega,t) \mapsto \mathcal{Q}_t(\omega)[0,t+a]$ is optional and the set

$$\{(t,\omega) \in [0,T) \times \Omega \ : \ \sigma_t(\omega) > t + \hat{\delta}\} = \cup_{n=1}^{\infty}\{(t,\omega) \ : \ \mathcal{Q}_t[0,t+\hat{\delta}+\frac{1}{n}] = 0\},$$

is also optional. This implies the measurability of

$$\Gamma = \{(t,\omega) \in [0,T) \times \Omega \ : \ \sigma_t(\omega) > t + \hat{\delta}, R_{t+\frac{1}{n}} > R_t \text{ for all } n\},$$

with respect to $\mathcal{B}([0,T)) \times \mathcal{F}_\infty$ where the filtration $\{\mathcal{F}_t\}$ is obtained by augmenting in the usual way the filtration $\sigma\{\mathcal{Q}_s, \nu_s, K_s, R_s \ : \ s \le t\}$. Therefore applying the Section Theorem for the measurable set Γ, we can find a $[0,T] \cup \{\infty\}$-valued random variable τ such that $\mathbb{P}(A_0^c) = \mathbb{P}(A_1 \cap A_2)$ for

$$A_1 = \{\tau < T \ : \ \sigma_\tau > \tau + \hat{\delta}\}, \quad \text{and} \quad A_2 = \{R_{\tau+\varepsilon} > R_\tau \text{ for all } \varepsilon > 0\}.$$

See [2, Lemma 5.9] for more details.

Therefore it is enough to show that on A_1, that is, if $\sigma_\tau(\omega) > \tau(\omega) + \hat{\delta}$ then there exists $\varepsilon = \varepsilon(\omega) > 0$ so that $R_{\tau+\varepsilon}(\omega) = R_\tau(\omega)$. This will prove $\mathbb{P}(A_0^c) = 0$ and therefore (6.9) holds a.s. We know that for any $a < b \in [0, T]$ we have

$$\bar{R}_b^N - \bar{R}_a^N \leq \bar{\mathcal{Q}}_a^N(a, b] + \bar{\alpha}_b^N(a, b] - \bar{\alpha}_a^N(a, b]. \tag{6.10}$$

The above follows from the fact that the total number of customers reneged in the interval $(a, b]$ must be smaller than the number of customers in the queue at time a with absolute deadlines in $(a, b]$ together with the total number of customers arrived in time interval $(a, b]$ with absolute deadline in $(a, b]$. Let $\varepsilon_J = J^{-1}\varepsilon$ for some $J \in \mathbb{N}$ and let $I_k, k = 1, \ldots, J$ denote the partition

$$I_k = (t_{k-1}, t_k], \quad t_k = \tau + k\varepsilon_J,$$

of $(\tau, \tau + \varepsilon]$. From (6.10)

$$\begin{aligned}
\bar{R}_{\tau+\varepsilon}^N - \bar{R}_\tau^N &= \sum_{k=1}^J (\bar{R}_{t_k}^N - \bar{R}_{t_{k-1}}^N) \\
&\leq \sum_{k=1}^J \Big(\bar{\alpha}_{t_k}^N(I_k) - \bar{\alpha}_{t_{k-1}}^N(I_k) + \bar{\mathcal{Q}}_{t_{k-1}}^N(I_k) \Big).
\end{aligned} \tag{6.11}$$

Since

$$\sup_{s\in[0,T]} d_{\mathcal{M}}(\bar{\mathcal{Q}}_s^N, \mathcal{Q}_s) \to 0, \quad \text{as} \quad N \to \infty, \text{ a.s.},$$

and $\mathcal{Q}$ does not have atoms we get

$$\sup_{s\in[0,T]} \sup_{x\in\mathbb{R}_+} |\bar{\mathcal{Q}}_s^N(x, \infty) - \mathcal{Q}_s(x, \infty)| \to 0, \quad \text{as } N \to \infty, \text{ a.s.} \tag{6.12}$$

On A_1 we have $\mathcal{Q}_\tau[0, \tau + \hat{\delta}] = 0$. Now suppose that Assumption 3.3 holds. Since (ν, K) satisfies (6.1) for every sample path, it also satisfies the conclusion of Lemma 3.2. Thus we can find δ and a compact set $\mathbf{C} \subset [0, H^s) \setminus \mathcal{Z}$ so that

$$\begin{aligned}
&\sup_{s\in[0,T]} \nu_s(\mathbf{C}^c) < \frac{1}{4}, \quad \sup_{s\in[0,T]} \lambda_s \pi[0, \delta] < \frac{3}{4} \inf_{x\in\mathbf{C}} h(x), \\
&G(\delta) < \frac{1}{10}\,.
\end{aligned}$$

Take any $\varepsilon \in (0, \hat{\delta} \wedge \delta)$. Then one can follow the proof of Lemma 4.4 to conclude that $\mathcal{Q}_s[0, \tau + \varepsilon] = 0$ for all $s \in [\tau, \tau + \varepsilon]$ almost surely on A_1. Now we deduce similar conclusion under Assumption 3.4. Let Assumption 3.4 hold. Since $\sigma_\tau > \tau + \hat{\delta}$ on A_1 we also have $\mathcal{Q}_\tau[0, \tau + \hat{\delta}] = 0$. Denote $\hat{\alpha}_s = \alpha_s - \alpha_\tau$. Similarly, define $\hat{K}_s = K_s - K_\tau$, $\hat{R}_s = R_s - R_\tau$. Then from (6.8) we obtain for $s \in [\tau, T]$ that for any $x \leq \tau + \hat{\delta}$

$$\begin{aligned}
\mathcal{Q}_s[0, x] &= \mathcal{Q}_\tau[0, x] + \hat{\alpha}_s[0, x] - \hat{K}_s - \hat{R}_s + \sup_{u\leq s}(\hat{K}_s + \hat{R}_s - \hat{\alpha}_s[0, x] - \mathcal{Q}_\tau[0, x])^+ \\
&= \hat{\alpha}_s[0, x] - \hat{K}_s - \hat{R}_s + \sup_{u\leq s}(\hat{K}_s + \hat{R}_s - \hat{\alpha}_s[0, x])^+.
\end{aligned}$$

Since $\pi[0,\kappa_1]=0$, if we choose $\varepsilon=\frac{\kappa_1}{2}\wedge\hat{\delta}$ we have for $s\le\tau+\hat{\delta}$ that

$$\hat{\alpha}_s[0,\tau+\varepsilon]=\int_\tau^s \lambda_u \mathbf{1}_{\{\tau+\varepsilon\ge u\}}\pi[0,\tau+\varepsilon-u]du=0.$$

Thus combining the above two displays we have $\mathcal{Q}_s[0,\tau+\varepsilon]=0$ for $s\in[\tau,\tau+\varepsilon]$. Thus using (6.12) we get

$$\sup_{s\in[\tau,\tau+\varepsilon]}\bar{\mathcal{Q}}^N_s[0,\tau+\varepsilon]\to 0,\quad \text{as } N\to\infty, \text{ a.s. on } A_1. \tag{6.13}$$

It should be noted that our choice of ε is non-random on A_1. Now we consider the summation

$$S^N_J=\sum_{k=1}^J\Big(\bar{\alpha}^N_{t_k}(I_k)-\bar{\alpha}^N_{t_{k-1}}(I_k)\Big).$$

For fixed J we see that as $N\to\infty$ we get $S^N_J\to S_J$ a.s., where

$$\begin{aligned}S_J&=\sum_{k=1}^J\Big(\alpha_{t_k}(I_k)-\alpha_{t_{k-1}}(I_k)\Big),\\&=\sum_{k=1}^J\int_{t_{k-1}}^{t_k}\mathbf{1}_{\{t_k\ge u\}}\lambda_s\pi[0,t_k-u]du\\&\le T\sup_{s\in[0,T]}\lambda_s\,\pi[0,\varepsilon_J],\end{aligned}$$

where the first equality is due to the fact that α does not have atoms. Therefore letting $N\to\infty$ in (6.11) and using (6.13), we have on A_1

$$R_{\tau+\varepsilon}-R_\tau\le T\sup_{s\in[0,T]}\lambda_s\,\pi[0,\varepsilon_J],$$

where J is arbitrary. Since $\lim_{x\to0}\pi[0,x]=0$ we get from above that $R_{\tau+\varepsilon}-R_\tau=0$ almost surely on A_1. This completes the proof. □

Acknowledgment. The authors are indebted to the anonymous referees for their careful review and suggestions. AB acknowledges the hospitality of the Department of Electrical Engineering in Technion while he was visiting at the early stages of this work. The research of RA was supported in part by grant 1184/16 of the Israel Science Foundation. The research of AB was supported in part by an INSPIRE faculty fellowship. The research of HK was supported in part by grant 764/13 of the Israel Science Foundation.

References

[1] S. Asmussen, *Applied Probability and Queues*. Springer (2008)

[2] R. Atar, A. Biswas, H. Kaspi, K. Ramanan, A Skorohod map on measure-valued paths with applications to priority queues. Ann. Appl. Probab. *To appear.*

[3] R. Atar, A. Biswas, H. Kaspi, Fluid limits of G/G/1+G queues under the non-preemptive earliest-deadline-first discipline. *Math. Oper. Res.* 40, No. 3, 683–702 (2015)

[4] R. Atar, H. Kaspi and N. Shimkin, Fluid limits for many-server systems with reneging under a priority policy. *Math. Oper. Res.*, Vol. 39, No. 3, 672–696 (2014).

[5] K. Burdzy, W. Kang and K. Ramanan, The Skorokhod problem in a time-dependent interval, *Stoch. Proc. Appl.* 119 (2009), no. 2, 428–452.

[6] L. Brown, G. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zetlyn and L. Zhao. Statistical analysis of a telephone call cemter: a queueing-science perspective. *J. Amer. Statist. Assoc.* 100, (2005) No. 469, 36–50.

[7] L. Decreusefond and P. Moyal, Fluid limit of a heavily loaded EDF queue with impatient customers, *Markov Proc. Rel. Fields* 14 (2008), 131–158,

[8] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes. Vol. I. Elementary theory and methods.* Second edition. Probability and its Applications (New York). Springer-Verlag, New York, 2003.

[9] B. Doytchinov, J. Lehoczky and S. Shreve, Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* 11 (2001), No. 2, 332–378.

[10] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence.* Wiley, 1986.

[11] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29 (1981), no. 3, 567–588

[12] J. Jacod and A.N. Shiryaev, *Limit Theorems for Stochastic Processes.* Springer-Verlag 1987.

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

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

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

[16] L. Kruk, Invariant states for fluid models of EDF networks: Nonlinear lifting map. *Probab. Math. Statist.*, 30 (2010): 289–315.

[17] L. Kruk, J. Lehoczky, K. Ramanan and S. Shreve, Heavy-traffic analysis for EDF queues with reneging, *Ann. Appl. Probab.* 21 (2011), no. 2, 484–545.

[18] A. Mandelbaum and P. Momćilović, Personalized queues: the customer view, via least-patience-first routing. Preprint.

[19] P. Moyal, On queues with impatience: stability, and the optimality of Earliest Deadline First, *Queueing Syst.* 75, 2–4, 211-242, 2014.

[20] S. S. Panwar and D. Towsley, On the optimality of the STE rule for multiple server queues that serve customer with deadlines. Technical Report 88–81, Dept. of Computer and Information Science, Univ. Massachusetts, Amherst.

[21] S. S. Panwar, D. Towsley and J. K. Wolf, Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service, *Journal of the Association for Computing Machinery*, Vol.35, pp.832–844, 1988.

[22] J. Reed. The G/GI/N queue in the Halfin-Whitt regime, *Ann. Appl. Probab.* 19 (2009), No. 6, 2211–2269.

[23] A. Walsh Zuñiga. Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions. *Stoch. Proc. Appl.* 124 (2014), no. 3, 1436–1468.

[24] W. Whitt. Fluid models for multiserver queues with abandonments, *Oper. Res.* 54 (2006) No. 1, 37–54.

[25] J. Zhang. Fluid models of many-server queues with abandonment. *Queueing Syst.* Vol 73 (2013), no. 2, 147–193.