Fluid limits for earliest-deadline-first networks

Rami Atar* Yonatan Shadmi*

March 15, 2021

Abstract

This paper analyzes fluid scale asymptotics of two models of generalized Jackson networks employing the earliest deadline first (EDF) policy. One applies the 'soft' EDF policy, where deadlines are used to determine priority but jobs do not renege, and the other implements 'hard' EDF, where jobs renege when deadlines expire, and deadlines are postponed with each migration to a new station. The arrival rates, deadline distribution and service capacity are allowed to fluctuate over time at the fluid scale. Earlier work on EDF network fluid limits, used as a tool to obtain stability of these networks, addressed only the soft version of the policy, and moreover did not contain a full fluid limit result. In this paper, tools that extend the notion of the measure-valued Skorokhod map are developed and used to establish for the first time fluid limits for both the soft and hard EDF network models.

AMS subject classifications: 60K25, 60G57, 68M20

Keywords: measure-valued Skorokhod map, measure-valued processes, fluid limits, earliest deadline first, generalized Jackson networks

1 Introduction

This work continues a line of research initiated in [4] and expanded in [3], in which a Skorokhod map (SM) is used in conjunction with stochastic evolution equations in measure space, to characterize scaling limits of queueing systems. Specifically, the measure-valued Skorokhod map (MVSM) of [4], an infinite-dimensional analogue of various well-known SMs in the finite-dimensional orthant, is a tool for analyzing policies that prioritize according to a continuous parameter. This notion has yielded new results on *fluid* or *law of large numbers* (LLN) scaling limits of queueing systems implementing the policies *earliest deadline first* (EDF), *shortest job first* (SJF) and *shortest remaining processing time* (SRPT) [4], as well as EDF in a many-server scaling [3]. In the case of EDF, the state of the system is given by a measure ξ on $\mathbb{R}$, where $\xi(B)$ expresses the queue length associated with all jobs in the buffer that have deadlines in the set B, for $B \subset \mathbb{R}$ a Borel set. The aforementioned continuous parameter corresponds to a job's deadline, and the MVSM encodes priority by enforcing the rule that, for every $x \in \mathbb{R}$, work with deadline $> x$ can be transferred from the buffer to the server only at times when $\xi((-\infty, x]) = 0$, that is, when there is no work in the buffer associated with deadline $\leq x$. Thus the MVSM of [4] is useful for the analysis of a *single node*. The goal of this paper is to address EDF in *network setting*. To this end two multidimensional counterparts of the MVSM are developed involving multiple measure-valued processes, one consisting of a

*Viterbi Faculty of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

map that combines properties of the MVSM and of the *SM with oblique reflection in the orthant* (SMOR), and another is where the MVSM serves as a building block in a system of equations.

EDF is a rational choice of service policy for systems in which job urgency is quantifiable. Its practical significance stems from its use in real time applications such as telecommunication networks [1], real time operating systems [9], and emergency department operations [18]. The policy has two main versions:

- soft EDF: all jobs are served whether their deadline expires or not; deadlines serve only to determine the priority of a job,
- hard EDF: deadlines determine priority and jobs whose deadlines expire renege the system.

The soft and hard versions of this policy are also known in the literature as EDF without and with reneging, respectively, and EDF is also referred to as *earliest due date first served*. A further subdivision is according to whether service is preemptive or non-preemptive; in this paper we restrict our attention to the non-preemptive disciplines. Although fluid limits of soft EDF queueing networks have been studied before in relation to the question of their stability in [8], [19], these papers have established only partial results as far as convergence was concerned (most significantly, for stability analysis only convergence along subsequences is required, and indeed only such convergence was proved; see more details below). In this paper, the two aforementioned counterparts of the MVSM yield for the first time fluid limits for both the soft and the hard versions of EDF networks.

There has been great interest in this policy in theoretical studies. It has been argued to possess optimality properties, and much effort has been devoted to studying its scaling limits [24], [14], [13], [21], [20], [22], [2], [6], mostly in the single node case. The hard EDF policy was shown to be optimal with respect to expected number of reneging jobs [27], for the $M/G/1+G$ and $G/D/1+G$ queues. It was also showed that if there exists an optimal policy for the $G/G/1+G$ queue then EDF is optimal. Optimality properties of EDF were further studied with respect to cumulative reneged work [22] as well as steady state probability of loss [26]. Fluid models of multi-class soft EDF Jackson networks were studied in [20], focusing on invariant states of the dynamical system and characterizing its invariant manifold.

As for scaling limits, most effort has been devoted to the heavy traffic regime. The paper [24] studied diffusion limits of soft EDF queues under heavy traffic assumptions as Markov processes in Banach spaces. An expression for the lead time profile given the queue length was obtained. An important line of research started from [14], a paper that pioneered (along with [15]) the use of measure-valued processes for scaling limits of queueing-related models. In this paper the diffusion limit of soft EDF queues was characterized. Further significant development was in [22], that used SM techniques to show convergence of a hard EDF queue to a diffusion limit. Specifically, the total workload process was shown to converge to a doubly reflected Brownian motion, and the limit of the underlying state process, that is again a measure-valued process, is given in terms of this reflected Brownian motion. The fact that the total workload in the system is sufficient to recover the complete state of the system is due to a certain *state space collapse* (SSC) phenomenon, which remarkably generalizes the well known SSC for priority queues. SSC, as a phenomenon and as a tool for proving convergence, is unique to the heavy traffic diffusion regime, and thus cannot be of help in the fluid limit regime that is of interest in this paper.

Hard EDF fluid limits were established in [13] and [2]. These papers considered the $M/M/1+G$ and, respectively, $G/G/1+G$ models. The paper [4] introduced the MVSM and used it to establish fluid limits for several models, including soft and hard EDF. The paper [3] employed this approach to analyze EDF in a many-server fluid limit regime.

Scaling limits of EDF in network setting were studied in [25] and [23] (in heavy traffic) and in [8] and [19] (at fluid scale). The paper [25] extended [24] to a network setting. Given the queue lengths at the

nodes, the Fourier transform of the lead time distribution was represented as the solution of a fixed point equation. The paper [23] extended [14] to multi-class acyclic networks.

Much closer to the subject of this paper are the results on fluid scale limits established in [8] and [19]. It is well known that fluid models are useful in proving stability of queueing networks, by reducing the problem of stability of the stochastic dynamical system representing the queueing network into the stability of all solutions to a deterministic dynamical system, usually represented in terms of deterministic evolution equations, comprising the fluid model. This approach, that was first formulated in general terms in [12], was taken in [8] to establish the stability of soft EDF networks, in fact in the broader, multi-class setting, provided that they are sub-critically loaded. The approach is based on showing stability properties of *any* solution to the fluid model equations, and does not require that uniqueness holds for these solutions (for a given initial condition). As a consequence, uniqueness of solutions to the fluid model equations is not required, nor is it established in [8]. A related issue is that when using this method it suffices to establish convergence of the rescaled processes along subsequences. Indeed, fluid limits are proved only along subsequences, and so the full convergence to a fluid limit is not established there.

It is also important to point out that the scaling in [8] is such that the gap between time of arrival and deadline vanishes in the limit. Consequently, the asymptotics are indistinguishable from that under a policy that prioritizes by order of arrival, namely *first in system first out* (FISFO). In contrast, under our scaling the gaps alluded to above remain of order one, and are therefore captured by the limiting dynamics. In this we follow the treatment in [4] and [3]. This aspect is also similar to the nature of the limit dynamics in most of the aforementioned papers on EDF in heavy traffic, where the limiting system's state is given by a nondegenerate measure that accounts for a variety of deadlines.

The stability problem was also studied via the fluid limit approach in [19] in the more complicated case of preemptive multiclass EDF networks, under the assumption that customers have fixed routes through the system. In [19] too it was not claimed, nor is it obvious, that the fluid model equations uniquely characterize the fluid model solutions for a given initial condition, and limits were only established along subsequences. The same paper also studied the stability of hard EDF networks (as well as several other network models with reneging), though not via fluid models, hence it did not address the problem of fluid limits of hard EDF networks.

Our first main result is the fluid limit for soft EDF Jackson networks. The approach is based on a tool that we develop, that combines properties of the aforementioned MVSM, which encodes the priority structure, and SMOR, which encodes the structure of the flow within the network. It is well known since [29] that the heavy traffic asymptotics of Jackson networks is given by a reflected Brownian motion in the orthant with oblique reflection vector field, a process that can be represented in terms of a corresponding SMOR. The same SMOR has been used since then to study Jackson networks in other asymptotic regimes, namely fluid limits [11, Ch. 7], and large deviations [5]. The way we use it in this paper is as follows. We represent the state of the system in terms of a *vector measure* (an $\mathbb{R}^d$ valued set function) ξ on $\mathbb{R}$, where for a Borel set $B \subset \mathbb{R}$, $\xi(B) = (\xi^i(B))$ expresses the workload associated with jobs having deadlines in B in the various buffers i. The resulting extended version of the MVSM, that we call the *vector measure valued SM* (VMVSM) is a transformation on the space of trajectories with values in the space of vector measures on $\mathbb{R}$. It has the property that for each deadline level $x \in \mathbb{R}_+$, the trajectory $t \mapsto \xi_t([0,x])$ is given as the image of the data of the problem (a suitable version of the so called netput process) under a SMOR. This provides a generalization of the concept from [4]. The convergence result is valid under very mild probabilistic assumptions. Note that fluid limits of soft EDF networks immediately imply fluid limits of FISFO networks, if the deadlines are set to be the arrival times.

Our second contribution is the treatment of hard EDF networks. The model is considerably more difficult as far as scaling limits are concerned, a fact reflected by the very small number of papers on the

subject. In particular, the method of [4] to handle hard EDF for a single node does not extend to networks (in [4] uniqueness for the fluid model solutions is obtained via pathwise minimality, a property that does not generalize to networks – counterexamples can be constructed). In our treatment we make two model assumptions that differ from the soft EDF model besides the obvious difference between the two policies.

First, the deadlines of a job are not fixed but increase whenever it migrates to a new service station. The problem appears to be considerably harder with fixed deadlines, having to do with regularity of the arrivals processes into each station. That is, in a model with fixed deadlines, when one of the nodes becomes supercritical, jobs begin to renege at some point, causing a potentially large number of jobs to arrive at other stations very close to their deadline. Consequently, small (asymptotically negligible) perturbations in deadlines may result in large (fluid scale) differences in the system's state. We leave open the important question of establishing fluid limits with fixed deadlines. Under the dynamic deadline assumption we are able to harness the single node results from [4] to the network setting. An important part of the proof is devoted to establishing regularity properties of arrival processes into each station, composed of exogenous and endogenous arrival processes (due to routing within the system), affected by the reneging in a nontrivial way. The aforementioned assumption allows us to use an argument by induction over squares in the time-deadline plane, where at each step the size of the square side is increased. This is used in treating both the fluid model equations and the convergence result.

Second, the service time distributions are assumed to have bounded support. This eases the notation for a certain technical reason related to the possibility that, while a job is in service, the (extended) deadline expires in the station it may be routed to. This assumption is not difficult to remove but we keep it because it does simplify the form of the model equations.

Finally we note that the techniques developed in this paper are likely to be useful for networks implementing other related disciplines, such as SJF.

The organization of this paper is as follows. At the end of this section we introduce notation used throughout. In §2 we describe the VMVSM and the soft EDF queueing network, and then state and prove the convergence result. In §3 we first describe the fluid model equations and prove uniqueness of solutions to these equations, then we formulate the queueing network model under hard EDF, state the main convergence result and provide its proof.

Notation

In what follows, $\mathbb{R}_+ = [0, \infty)$. For a Polish space $\mathcal{S}$, $\mathbb{C}_\mathcal{S}$ and $\mathbb{D}_\mathcal{S}$ denote the space of continuous and, respectively, cádlág functions from $\mathbb{R}_+$ into $\mathcal{S}$; if $\mathcal{S} = \mathbb{R}$ we simply write $\mathbb{C}$ and $\mathbb{D}$. Denote by $\mathbb{D}_+$, respectively $\mathbb{D}^\uparrow$, the subset of $\mathbb{D}$ of non-negative functions, respectively of non-decreasing and non-negative functions, and apply a similar notation to $\mathbb{C}$. We also define

$$\mathbb{D}_0^K = \left\{ f \in \mathbb{D}^K : \quad f^i(0) \geq 0 \text{ for all } 1 \leq i \leq K \right\}.$$

$\mathcal{M}$ denotes the space of finite Borel measures on $\mathbb{R}_+$, endowed with the topology of weak convergence. Its subset of atomless measures is denoted by $\mathcal{M}_\sim$. It is well known that the topology of weak convergence is metrized by the Levy-Prohorov metric, denoted $d_\mathcal{L}$. Define the subset of $\mathbb{D}_\mathcal{M}$ of non-decreasing elements as $\mathbb{D}_\mathcal{M}^\uparrow = \{\zeta \in \mathbb{D}_\mathcal{M} \colon t \mapsto \int_{\mathbb{R}_+} f(x)\zeta_t(dx)$ is non-decreasing for any continuous, bounded, non-negative function $f\}$. With a slight abuse of notation we denote $\mathbb{D}^{\uparrow K} = \left(\mathbb{D}^\uparrow\right)^K$ and $\mathbb{D}_\mathcal{M}^{\uparrow K} = \left(\mathbb{D}_\mathcal{M}^\uparrow\right)^K$. The support of a measure $\zeta \in \mathcal{M}$ is denoted by $\mathrm{supp}\,[\zeta]$. For $\zeta \in \mathcal{M}$ we write $\zeta[a,b]$ for $\zeta([a,b])$, and similarly for $[a,b)$, etc. We use the convention that for $a > b$, $[a,b] = \emptyset$. $\mathbb{D}$ and $\mathbb{D}_\mathcal{M}$ are equipped with the corresponding J_1 Skorokhod topologies, and $\mathbb{D}^K$ and $\mathbb{D}_\mathcal{M}^K$ are equipped with the product topologies. For $x \in \mathbb{R}^K$, x^i denotes

the i-th component, and $\|x\| = \max_{1\le i\le K} \left|x^i\right|$. For $x \in \mathbb{D}^K$ denote $\|x\|_T = \sup_{t\in[0,T]} \|x(t)\|$. The modulus of continuity of a function $f : \mathbb{R}_+ \to \mathbb{R}$ is denoted by

$$w_T(f,\epsilon) = \sup\{|f(s) - f(t)| : s,t \in [0,T], |s-t| \le \epsilon\}.$$

2 Soft EDF networks

In this section we develop the VMVSM, which combines elements of the MVSM and the SMOR. We then use it to represent and establish the fluid limit of soft EDF networks. We begin by introducing, in §2.1.1, the MVSM, and then a Skorokhod problem whose well posedness is stated and proved. This gives rise to the VMVSM. Then, in §2.1.2, the queueing model and its rescaling are introduced. We then state the main result. Its proof appears in §2.2.

2.1 Model and main results

2.1.1 A Skorohod problem in measure space

The measure valued Skorokhod problem (MVSP) introduced in [4] is as follows.

Definition 1 (MVSP)**.** Let $(\alpha,\mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}$. Then $(\xi,\beta,\iota) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}$ is said to solve the MVSP for the data (α,μ) if, for each $x \in \mathbb{R}_+$,

1. $\xi[0,x] = \alpha[0,x] - \mu + \beta(x,\infty) + \iota$,
2. $\int_{[0,\infty)} \xi_s[0,x]\, d\beta_s(x,\infty) = 0$,
3. $\int_{[0,\infty)} \xi_s[0,x]\, d\iota(s) = 0$,
4. $\beta[0,\infty) + \iota = \mu$,

where in 2. and 3. the integration variable is s.

It was shown that there exists a unique solution to the MVSP ([4, Proposition 2.8]), and thus, the MVSP defines a map $(\alpha,\mu) \mapsto (\xi,\beta,\iota)$. Moreover, this map has certain continuity properties that were key in establishing scaling limits for both soft and hard EDF [4, Theorem 5.4]. The approach that we take here builds on these ideas but replaces ξ and β by vector measures. We motivate our definitions by describing a fluid model for an EDF network.

An informal description is followed by a precise mathematical formulation. Consider a queueing network with K nodes; each node consists of a server and a queue, that accommodates fluid. Let $P \in \mathbb{R}^{K\times K}$ be a given substochastic matrix. Exogenous arrivals form an input to the system. They consist of fluid entering the various queues. Each unit of arriving mass has an associated distribution of deadlines. The servers prioritize the mass with the smallest deadline. After service, the mass splits between the nodes according to P. That is, the stream leaving server i splits so that a fraction P_{ij} routes to queue j. These streams form the endogenous arrival processes. The remaining fraction, $1 - \sum_{j=1}^K P_{ij}$, exits the system. The deadlines do not vary during the entire sojourn in the system.

Define for the i-th queue, respectively, the exogenous arrival process, cumulative potential effort, queue content, and departure process, as follows:

$$\begin{aligned}
\alpha^i \in \mathbb{D}^{\uparrow}_{\mathcal{M}}: &\quad \alpha^i_t[0,x] \text{ is the mass to have exogenously entered the } i\text{-th queue by time } t \\
&\quad \text{with deadlines in } [0,x], \\
\mu^i \in \mathbb{D}^{\uparrow} &: \quad \mu^i(t) \text{ is the total mass the } i\text{-th server can serve by time } t, \text{ if it is never idle}, \\
\xi^i \in \mathbb{D}_{\mathcal{M}}: &\quad \xi^i_t[0,x] \text{ is the mass in the } i\text{-th queue at time } t \text{ with deadlines in } [0,x], \\
\beta^i \in \mathbb{D}^{\uparrow}_{\mathcal{M}}: &\quad \beta^i_t[0,x] \text{ is the mass to have left the } i\text{-th queue by time } t \text{ with deadlines in } [0,x].
\end{aligned} \tag{1}$$

The initial condition of the buffer content is already contained in α, namely it is given by α_0. Let the idleness (or lost effort) process be defined by $\iota^i = \mu^i - \beta^i[0,\infty)$. The following balance equation holds:

$$\xi^i_t[0,x] = \alpha^i_t[0,x] + \sum_{j=1}^{K} P_{ji}\beta^j_t[0,x] - \beta^i_t[0,x]. \tag{2}$$

The work conservation property and the priority structure of EDF are expressed through

$$\begin{aligned}
\int \xi^i_t[0,x]\, d\iota^i(t) = 0, \quad 1 \le i \le K, \\
\int \xi^i_t[0,x]\, d\beta^i_t(x,\infty) = 0, \quad 1 \le i \le K.
\end{aligned} \tag{3}$$

One can recognize similarities to the MVSP, although the objects in our equations are vector-valued parallelrs of those from the MVSP, and moreover, the equations are coupled. This motivates us to define a multi-dimensional version of the MVSP.

Definition 2 (VMVSP)**.** Let $P \in \mathbb{R}^{K\times K}$ be a substochastic matrix and let $R = I - P^T$. Let $(\alpha,\mu) \in \mathbb{D}^{\uparrow K}_{\mathcal{M}} \times \mathbb{D}^{\uparrow K}$. Then $(\xi,\beta,\iota) \in \mathbb{D}^{K}_{\mathcal{M}} \times \mathbb{D}^{\uparrow K}_{\mathcal{M}} \times \mathbb{D}^{\uparrow K}$ is said to solve the VMVSP associated with the matrix P for the data (α,μ) if, for each $x \in \mathbb{R}_+$,

1. $\xi[0,x] = \alpha[0,x] - R\beta[0,x]$,
2. $\int_{[0,\infty)} \xi^i_s[0,x]\, d\beta^i_s(x,\infty) = 0$ for $i \in \{1,...,K\}$,
3. $\int_{[0,\infty)} \xi^i_s[0,x]\, d\iota^i(s) = 0$ for $i \in \{1,...,K\}$,
4. $\beta[0,\infty) + \iota = \mu$.

Note that the motivating fluid model indeed requires that ξ_t be a nonnegative vector measure for each t, and that $t \mapsto \beta_t(x,\infty) + \iota(t)$ be non-decreasing and non-negative. These properties are assured by the above notion on VMVSM.

A square matrix is called an M-matrix if it has positive diagonal elements, non-positive off-diagonal elements, and a non-negative inverse. By [11, Lemma 7.1], for a non-negative matrix G whose spectral radius is strictly less than 1, also called convergent, $I - G$ is an M-matrix.

Theorem 1. *Let $P \in \mathbb{R}^{K\times K}$ be a convergent substochastic matrix, and $(\alpha,\mu) \in \mathbb{D}^{\uparrow K}_{\mathcal{M}} \times \mathbb{D}^{\uparrow K}$ be such that $\mu(0) = 0$. Then the VMVSP associated with the matrix P and the data (α,μ) has a unique solution.*

Proof. The uniqueness argument is based on SM theory for oblique reflection in the orthant. As for existence, our strategy is to construct a candidate and show that it is a solution, relying in a crucial way on a monotonicity result due to Ramasubramanian.

We first present the existence argument. To construct a candidate, first rewrite the first condition of the VMVSP as

$$\xi_t[0,x] = \alpha_t[0,x] - R\mu(t) + R(\beta_t(x,\infty) + \iota(t)). \tag{4}$$

The *oblique reflection mapping theorem* (ORMT), [16], [11, Theorem 7.2] states that for an M-matrix R, for every $u \in \mathbb{D}_0^K$ there exists a unique pair $(z,y) \in \mathbb{D}_+^K \times \mathbb{D}^{\uparrow K}$ satisfying

$$z = u + Ry \tag{5}$$

$$\int_{[0,\infty)} z^i(s)\, dy^i(s) = 0, \quad 1 \le i \le K. \tag{6}$$

The solution map, denoted $\Gamma : \mathbb{D}_0^K \to \mathbb{D}_+^K \times \mathbb{D}^{\uparrow K}$, is thus uniquely defined by the relation: $(z,y) = \Gamma(u) = (\Gamma_1(u), \Gamma_2(u))$ iff (5)–(6) hold. It is well known that the maps Γ_1 and Γ_2 are Lipschitz from $\mathbb{D}_0^K$ to $\mathbb{D}_+^{\uparrow K}$ and $\mathbb{D}_+^K$ in the sense that for any $T > 0$

$$\|\Gamma_i(u_1) - \Gamma_i(u_2)\|_T \le L\|u_1 - u_2\|_T, \quad i = 1,2, \tag{7}$$

where the constant L depends only on R.

The matrix P^T is non-negative and convergent by assumption, hence R is an M-matrix. Moreover, $\alpha_0[0,x] - R\mu(0) = \alpha_0[0,x] \ge 0$. Also ,if (ξ,β,ι) is a solution, then $(\xi[0,x], \beta(x,\infty) + \iota) \in \mathbb{D}_+^K \times \mathbb{D}^{\uparrow K}$. Finally, Equation (4) has the form of (5), and conditions 2 and 3 of the VMVSP correspond to (6). Therefore, the conditions of the ORMT are satisfied, so we can construct a candidate as follows. Define

$$\mathring{\xi}(x) = \Gamma_1(\alpha[0,x] - R\mu) \in \mathbb{D}_+^K, \qquad \mathring{\beta}(x) + \iota = \Gamma_2(\alpha[0,x] - R\mu) \in \mathbb{D}^{\uparrow K}. \tag{8}$$

The path ι can be recovered by taking x to infinity, yielding $\iota = \Gamma_2(\alpha[0,\infty) - R\mu)$. One can then find $\mathring{\beta}_t(x)$ by $\mathring{\beta}(x) = \Gamma_2(\alpha[0,x] - R\mu) - \iota = \Gamma_2(\alpha[0,x] - R(\mu - \iota))$.

It is now argued that these functions define vector measure valued paths via the relations $(\mathring{\xi}_t(x), \mathring{\beta}_t(x)) = (\xi_t[0,x], \beta_t(x,\infty))$. We must show that $\mathring{\xi}_t(x)$ is right-continuous and non-decreasing in x, $\mathring{\beta}(x) - \beta(y)$ is in $\mathbb{D}^{\uparrow K}$ for $x < y$ and that $\mathring{\beta}(x)$ right-continuous in x. The right-continuity follows by the Lipschitz property (7) and Lemma 2.4 in [4]. Next, Theorem 4.1 of [28] states that the map Γ is monotone in the following sense. Let $u_1, u_2 \in \mathbb{D}_0^K$ such that $u_2 - u_1 \in \mathbb{D}^{\uparrow K}$, and let $(z_i, y_i) = \Gamma(u_i), i = 1,2$. Then $z_2 - z_1 \in \mathbb{D}_+^K$ and $y_1 - y_2 \in \mathbb{D}^{\uparrow K}$. Using this theorem in the setting of the VMVSP, let $x < y$ and denote $u_1 = \alpha[0,x] - R\mu$, and $u_2 = \alpha[0,y] - R\mu$. Then $u_2 - u_1 = \alpha(x,y] \in \mathbb{D}^{\uparrow K}$. Therefore $\mathring{\xi}(y) - \mathring{\xi}(x) \in \mathbb{D}_+^K$ and $\mathring{\beta}(x) - \mathring{\beta}(y) \in \mathbb{D}^{\uparrow K}$. Thus

$$\xi[0,x] = \mathring{\xi}(x) = \Gamma_1(\alpha[0,x] - R\mu), \qquad \beta(x,\infty) = \mathring{\beta}(x) = \Gamma_2(\alpha[0,x] - R\mu) - \iota, \tag{9}$$

define two vector measure valued paths in $\mathbb{D}_{\mathcal{M}}^K$ and $\mathbb{D}_{\mathcal{M}}^{\uparrow K}$ respectively. This demonstrates existence.

As for uniqueness, it follows from the ORMT that any solution to the VMVSP must satisfy

$$(\xi[0,x], \beta(x,\infty) + \iota) = \Gamma(\alpha[0,x] - R\mu),$$

a relation that defines uniquely ξ, β and ι. □

Theorem 1 defines a map from $\mathbb{D}_{\mathcal{M}}^{\uparrow K} \times \mathbb{D}^{\uparrow K} \to \mathbb{D}_{\mathcal{M}}^K \times \mathbb{D}_{\mathcal{M}}^{\uparrow K} \times \mathbb{D}^{\uparrow K}$, namely the solution map of the VMVSM. We denote it by Θ. The dependence on P is not indicated explicitly. This notion is an extension of the MVSM from [4].

2.1.2 Queueing model and scaling

The queueing model is defined analogously to the fluid model. It is indexed by $N \in \mathbb{N}$, and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It consists of $K \geq 1$ service stations, each containing a buffer with infinite room and a server. The servers prioritize according to the EDF policy, and according to arrival times in case of a tie. Throughout this paper, for any parameter, random variable or process associated with the N-th system, say a^N, we use the notation $\bar{a}^N = N^{-1}a^N$ for normalization. A substochastic matrix $P \in \mathbb{R}^{K\times K}$, referred to as the routing matrix, is given, where the entry P_{ij} represents the probability that a job that completes service at the i-th server is routed to the j-th queue. Denote $R = I - P^T$. For each N, processes denoted by $\alpha^N = (\alpha^{i,N})$, $\mu^N = (\mu^{i,N})$, $\xi^N = (\xi^{i,N})$, and $\beta^N = (\beta^{i,N})$ are associated with the N-th system, which represent discrete versions of their fluid model counterparts. Specifically, for a Borel set $B \subset \mathbb{R}_+$, $\alpha_t^{i,N}(B)$ denotes the number of external arrivals into queue i up to time t with deadline in B, $\xi_t^{i,N}(B)$ denotes the number of jobs in queue i at time t with deadline in B, and $\beta_t^{i,N}(B)$ denotes the number of jobs with deadline in B transferred by time t from the i-th queue to the corresponding server. Indeed, in the queueing model, the job being served is not counted in the queue, and therefore there is a distinction between the number of jobs transferred from queue i and the number of jobs transferred from server i. Thus in addition to $\beta^{i,N}$ we introduce a process γ^N. Namely, $\gamma_t^{ij,N}(B)$ denotes the number of jobs with deadline in B that were transferred from server i to queue j by time t, where $j = 0$ corresponds to jobs leaving the system. For each i, $\gamma_t^{i,N}(B) := \sum_{j=0}^K \gamma_t^{ij,N}(B)$ gives the total number of jobs with deadline in B that departed server i by time t. The process $\mu^{i,N}(t)$ represents the cumulative service capacity of server i by time t. Let the busyness at time t, $B_t^{i,N}(B)$, be defined as the indicator of the event that at this time, a job with deadline in B occupies the i-th server, and let $B_{0-}^{i,N}(B)$ be its initial condition. Denote the total number of departures from server i by $D^{i,N}(t) := \gamma_t^{i,N}[0,\infty)$. Denote by $\xi_{0-}^{i,N}(B)$ the number of jobs present in the i-th queue just prior to time $t = 0$ with deadlines in B.

Next, define the cumulative effort and cumulative lost effort processes, respectively, as

$$T^{i,N}(t) = \int_{[0,t]} B_s^{i,N}[0,\infty)\, d\mu^{i,N}(s), \tag{10}$$

$$\iota^{i,N}(t) = \mu^{i,N}(t) - T^{i,N}(t). \tag{11}$$

To model the stochasticity of the service times a counting process $S^i(t)$ associated with each server i is introduced, assumed to be a renewal process for which the interarrival times have unit mean, with the convention $S^i(0) = 1$. Then the departure process is given by

$$D^{i,N}(t) = \gamma_t^{i,N}[0,\infty) = S^i\left(T^{i,N}(t)\right) - 1. \tag{12}$$

The various relations between these processes are described in what follows. The relation between γ^N, β^N and B^N is given by

$$B_{0-}^{i,N}[0,x] + \beta_t^{i,N}[0,x] = \gamma_t^{i,N}[0,x] + B_t^{i,N}[0,x]. \tag{13}$$

The balance equation for the queue content is

$$\xi_t^{i,N}[0,x] = \alpha_t^{i,N}[0,x] + \sum_{j=1}^K \gamma_t^{ji,N}[0,x] - \beta_t^{i,N}[0,x]. \tag{14}$$

In addition, the work conservation condition and EDF policy are expressed through

$$\int_{[0,\infty)} \xi_t^{i,N}[0,x]\, d\iota^{i,N}(t) = 0, \qquad 1 \le i \le K, \tag{15}$$

$$\int_{[0,\infty)} \xi_t^{i,N}[0,x]\, d\beta_t^{i,N}(x,\infty) = 0, \quad 1 \le i \le K. \tag{16}$$

For the routing process of server i, consider i.i.d. $\{0,1,\ldots,K\}$-valued random variables $\pi^{i,N}(n)$ with distribution given by

$$\mathbb{P}\left(\pi^{i,N}(n) = j\right) = \begin{cases} 1 - \sum_{k=1}^{K} P_{ik} & \text{if } j = 0, \\ P_{ij} & \text{if } j \in \{1,...,K\}. \end{cases}$$

The stochastic primitives $\{\pi^{i,N}(n)\}$, $\{S\}$ and $\{\alpha^N\}$ are assumed to be mutually independent. Moreover, $\pi^{i,N}(n)$ are assumed to be mutually independent across i, and so are S^i and $\alpha^{i,N}$.

Define for $j \in \{1,...,K\}$: $\theta^{ij,N}(n) = \mathbb{1}_{\{\pi^{i,N}(n)=j\}}$. The n-th job to depart server i is routed to server $\pi^{i,N}(n)$, or, if this random variable is zero, leaves the system. Thus, the n-th job is routed to server j if and only if $\theta^{ij,N}(n) = 1$. Let the jump times of $D^{i,N}$ be denoted by

$$\tau_n^{i,N} = \inf\{t \ge 0 : D^{i,N}(t) \ge n\}, \tag{17}$$

and let

$$\hat{\theta}^{ij,N}(t) = \sum_{n=1}^{\infty} \mathbb{1}_{\left[\tau_n^{i,N}, \tau_{n+1}^{i,N}\right)}(t)\, \theta^{ij,N}(n). \tag{18}$$

Then $\gamma^{ij,N}$ can be obtained from $\gamma^{i,N}$ by

$$\gamma_t^{ij,N}[0,x] = \int_{[0,t]} \hat{\theta}^{ij,N}(s)\, d\gamma_s^{i,N}[0,x]. \tag{19}$$

This completes the definition of the model.

Our main result concerning soft EDF networks is the following.

Theorem 2. *Assume that the routing matrix P is a convergent substochastic matrix, and let $(\alpha,\mu) \in \mathbb{C}_{\mathcal{M}}^{\uparrow K} \times \mathbb{C}^{\uparrow K}$ be such that $\mu(0) = 0$. Assume, moreover, that $\alpha_t \in \mathcal{M}_\sim$ for all t, and that there exists a constant C such that $\mathbb{E}\left[\sum_{i=1}^{K} \alpha_t^{i,N}[0,\infty)\right] \le CNt$. Finally, assume $(\bar{\alpha}^N, \bar{\mu}^N) \Rightarrow (\alpha,\mu)$. Then $\left(\bar{\xi}^N, \bar{\beta}^N, \bar{\iota}^N\right) \Rightarrow (\xi,\beta,\iota)$, where (ξ,β,ι) is the unique solution of the VMVSP with primitives (α,μ).*

The proof is given in the next section.

2.2 Proof

Recall the ORMT. To use this theorem, we wish to bring equations (14), (15) and (16) to a form compatible with the conditions (5) and (6). To this end, first define the error processes

$$\begin{aligned} e^{i,N}(t) &= \beta_t^{i,N}[0,\infty) + \iota^{i,N}(t) - \mu^{i,N}(t), \quad 1 \le i \le K, \\ E^{ij,N}(t,x) &= \gamma_t^{ij,N}[0,x] - P_{ij}\gamma_t^{i,N}[0,x], \quad 1 \le i,j \le K, x \in \mathbb{R}_+. \end{aligned} \tag{20}$$

Using (14) and normalizing,

$$\bar{\xi}_t^{i,N}[0,x] = \bar{\alpha}_t^{ij,N}[0,x] + \sum_{j=1}^{K} \bar{E}^{ji,N}(t,x) + \sum_{j=1}^{K} P_{ji}\left(\bar{\mu}^{j,N}(t) + \bar{e}^{j,N}(t)\right) - \bar{\mu}^{i,N}(t) - \bar{e}^{i,N}(t)$$
$$- \sum_{j=1}^{K} P_{ji}\left(\bar{B}_t^{j,N}[0,x] - \bar{B}_{0-}^{j,N}[0,x]\right) + \bar{\beta}_t^{i,N}(x,\infty) + \bar{\iota}^{i,N}(t) - \sum_{j=1}^{K} P_{ji}\left(\bar{\beta}_t^{j,N}(x,\infty) + \bar{\iota}^{j,N}(t)\right).$$

In vector notation, with $\mathbf{1} = (1,\ldots,1)^T \in \mathbb{R}^K$,

$$\begin{aligned}\bar{\xi}_t^N[0,x] =& \bar{\alpha}_t^N[0,x] + \bar{E}^{N,T}(t,x)\,\mathbf{1} - R\left(\bar{\mu}^N(t) + \bar{e}^N(t)\right) \\ &- P^T\left(\bar{B}_t^N[0,x] - \bar{B}_{0-}^N[0,x]\right) + R\left(\bar{\beta}_t^N(x,\infty) + \bar{\iota}^N(t)\right).\end{aligned} \tag{21}$$

The matrix R is an M-matrix as already argued, $\bar{\xi}^N$ is non-negative, and $\bar{\beta}_t^N(x,\infty) + \bar{\iota}^N$ is non-decreasing and non-negative. We can now invoke the ORMT, yielding for every x,

$$\begin{aligned}&\bar{\xi}^N[0,x] = \Gamma_1\left(\bar{\alpha}^N[0,x] + \bar{E}^{N,T}(\cdot,x)\,\mathbf{1} - R\left(\bar{\mu}^N + \bar{e}^N\right) - P^T\bar{B}^N[0,x] + P^T\bar{B}_{0-}^N[0,x]\right), \\ &\bar{\beta}^N(x,\infty) + \bar{\iota}^N = \Gamma_2\left(\bar{\alpha}^N[0,x] + \bar{E}^{N,T}(\cdot,x)\,\mathbf{1} - R\left(\bar{\mu}^N + \bar{e}^N\right) - P^T\bar{B}^N[0,x] + P^T\bar{B}_{0-}^N[0,x]\right).\end{aligned} \tag{22}$$

Lemma 1. *Let $E^{ij,N}(t,x)$ and $D^{i,N}(t)$ be as in (20) and (12). Then, for all i,j,N,t,x,*

$$\mathbb{E}\left[E^{ij,N}(t,x)^2\right] \le 41\mathbb{E}\left[D^{i,N}(t)\right].$$

Proof. We have

$$\begin{aligned}E^{ij,N}(t,x) &= \gamma_t^{ij,N}[0,x] - P_{ij}\gamma_t^{i,N}[0,x] \\ &= \int_{[0,t]}\left(\hat{\theta}^{ij,N}(s) - P_{ij}\right)d\gamma_s^{i,N}[0,x] \\ &= \sum_{n=1}^{D^{i,N}(t)}\left(\theta^{ij,N}(n) - P_{ij}\right)\left(\gamma^{i,N}_{\tau_n^{i,N}}[0,x] - \gamma^{i,N}_{\tau_{n-1}^{i,N}}[0,x]\right).\end{aligned}$$

To simplify notation in this proof, we omit in what follows the dependence on i,j and N. Denote

$$M_k(x) = \sum_{n=1}^{k}\left(\theta(n) - P\right)\left(\gamma_{\tau_n}[0,x] - \gamma_{\tau_{n-1}}[0,x]\right).$$

We show that $M_k(x)$ is a martingale on the filtration

$$\mathcal{F}_k = \sigma\left(\theta(n) : 1 \le n \le k,\ \gamma_{\tau_n}[0,y] : 1 \le n \le k+1, y \ge 0,\ \tau_n : 1 \le n \le k+1\right).$$

We first argue that $\theta(k+1)$ is independent of $\mathcal{F}_k$. By assumption, $\theta(k+1)$ is independent of

$$\mathcal{G}_k := \left(\theta(n), S(t), \alpha_t[0,y], \quad t,y \ge 0, 1 \le n \le k\right).$$

By construction, $\mathcal{G}_k \supset \mathcal{F}_k$. Therefore, $\theta(k+1)$ is independent of $\mathcal{F}_k$. The process $M_k(x)$ is adapted to $\{\mathcal{F}_k\}_{k\ge1}$, it satisfies $\mathbb{E}\left[|M_k(x)|\right] \le k < \infty$, and

$$\begin{aligned}&\mathbb{E}\left[\sum_{n=1}^{k+1}\left(\theta(n) - P\right)\left(\gamma_{\tau_n}[0,x] - \gamma_{\tau_{n-1}}[0,x]\right)\middle|\mathcal{F}_k\right] \\ &= M_k(x) + \left(\gamma_{\tau_{k+1}}[0,x] - \gamma_{\tau_k}[0,x]\right)\mathbb{E}\left[\theta(k+1) - P|\mathcal{F}_k\right] \\ &= M_k(x).\end{aligned}$$

Consequently it is a martingale.

Next, $D(t)$ is a stopping time with respect to $\{\mathcal{F}_k\}_{k\geq 1}$, because $\{D(t)\leq k\}=\{\tau_{k+1}>t\}$ and τ_{k+1} is $\mathcal{F}_k$-measurable. Therefore, $M_{k\wedge D(t)}$ is also a martingale, and by Fatou's lemma and Burkholder's inequality (see [17], Theorem 2.10)

$$\begin{aligned}\mathbb{E}\left[E^2(t,x)\right] &= \mathbb{E}\left[M^2_{D(t)}(x)\right]\\ &\leq \liminf_{k\to\infty}\mathbb{E}\left[M^2_{k\wedge D(t)}(x)\right]\\ &\leq 41\liminf_{k\to\infty}\mathbb{E}\left[\sum_{n=1}^{k\wedge D(t)}(\theta(n)-P)^2\left(\gamma_{\tau_n}[0,x]-\gamma_{\tau_{n-1}}[0,x]\right)^2\right]\\ &\leq 41\mathbb{E}\left[D(t)\right].\end{aligned}$$

□

Proof of Theorem 2. We write c for a generic positive constant whose value may change from line to line.

We first show that $\bar{\xi}^N\to\xi$ in probability. If $\alpha_1,\alpha_2\in\mathcal{M}$ then $d_{\mathcal{L}}(\alpha_1,\alpha_2)\leq\sup_x|\alpha_1[0,x]-\alpha_2[0,x]|$. Consequently it suffices to show for every T

$$\sup_x\left\|\bar{\xi}^N[0,x]-\xi[0,x]\right\|_T\to 0 \text{ in probability.} \tag{23}$$

By (7), (9) and (22),

$$\begin{aligned}&\sup_x\left\|\bar{\xi}^N[0,x]-\xi[0,x]\right\|_T\\ &\leq L\sup_x\left\|\bar{\alpha}^N[0,x]-\alpha[0,x]\right\|_T+L\sup_x\left\|\bar{E}^{N,T}(\cdot,x)\mathbf{1}\right\|_T+L\sup_x\left\|P^T\bar{B}^N[0,x]+P^T\bar{B}^N_{0-}[0,x]\right\|_T\\ &\quad+L\left\|R\left(\mu-\bar{\mu}^N\right)\right\|_T+L\left\|R\bar{e}^N\right\|_T.\end{aligned}$$

We now show that each of the terms converges in probability to 0.

• It follows directly from the definition of the Lévy metric that for $a_1,a_2\in\mathcal{M}$,

$$\sup_{x\in\mathbb{R}_+}|a_1[0,x]-a_2[0,x]|\leq d_{\mathcal{L}}(a_1,a_2)+w(a_2[0,\cdot],2d_{\mathcal{L}}(a_1,a_2)). \tag{24}$$

The fact that $\alpha^i\in\mathbb{D}^{\uparrow}_{\mathcal{M}}$ implies that for any $t\leq T$, $w(\alpha_t[0,\cdot],\delta)\leq w(\alpha_T[0,\cdot],\delta)$. Since $\alpha_T\in\mathcal{M}$ does not have atoms by assumption, we have $w(\alpha_T[0,\cdot],\epsilon)\to 0$ as $\epsilon\to 0$. Denote

$$d_{\mathcal{L},T}(\zeta_1,\zeta_2)=\sup_{t\in[0,T]}\max_{1\leq i\leq K}d_{\mathcal{L}}\left(\zeta^i_{1,t},\zeta^i_{2,t}\right),$$

for $\zeta_1,\zeta_2\in\mathbb{D}^K_{\mathcal{M}}$. By the convergence $\alpha^N\Rightarrow\alpha$ and the continuity of $t\mapsto\alpha_t$, it follows that $d_{\mathcal{L},T}\left(\alpha^N,\alpha\right)\to 0$ in probability. Hence by (24),

$$\sup_x\left\|\bar{\alpha}^N[0,x]-\alpha[0,x]\right\|_T\to 0 \text{ in probability.}$$

• For the second term, it suffices to show

$$\sup_x\sup_{t\in[0,T]}\max_{1\leq i,j\leq K}\left|\bar{E}^{ji,N}(t,x)\right|\to 0 \text{ in probability.}$$

To this end note that Lemma 1 implies

$$\mathbb{E}\left[\left|\bar{E}^{ij,N}(t,x)\right|^2\right] \le 41N^{-2}\mathbb{E}\left[D^{i,N}(t)\right] \le 41N^{-2}\mathbb{E}\left[\sum_{i=1}^{K}\alpha_T^{i,N}[0,\infty)\right] \le 41N^{-1}CT \to 0.$$

• The third term on the RHS is bounded by $2LK/N \to 0$.
• Next, $\left\|R\left(\mu-\bar{\mu}^N\right)\right\|_T \le c\left\|\mu-\bar{\mu}^N\right\|_T \to 0$ in probability by the assumed convergence of $\bar{\mu}^N$ to μ and continuity of μ.
• Finally, by the definition of $\bar{e}^N$,

$$\begin{aligned}\bar{e}^{i,N}(t) &= \bar{\beta}_t^{i,N}[0,\infty) + \bar{\iota}^{i,N}(t) - \bar{\mu}^{i,N}(t)\\ &= -\frac{1}{N} + \frac{B^{i,N}[0,\infty)}{N} + \frac{S^i\left(N\bar{T}^{i,N}(t)\right) - N\bar{T}^{i,N}(t)}{N}.\end{aligned}$$

Now, by the law of large numbers for renewal processes, for any constant c, $\sup_{t\in[0,c]}\left|\frac{S^i(Nt)-Nt}{N}\right| \to 0$ in probability. Moreover, for fixed T, $\bar{T}^{i,N}(T)$ are tight. It follows that

$$\sup_{t\in[0,T]}\left|\frac{S^i\left(N\bar{T}^{i,N}(t)\right) - N\bar{T}^{i,N}(t)}{N}\right| \to 0$$

in probability. Consequently, $\|\bar{e}^N\|_T \to 0$ in probability. This completes the proof of (23). We conclude that $\bar{\xi}^N \to \xi$.

The convergence of $\bar{\iota}^N$ and $\bar{\beta}^N$ to ι and β is shown similarly, by virtue of the Lipschitz property of Γ_2. We omit the details. This shows $(\bar{\xi}^N, \bar{\beta}^n, \bar{\iota}^N) \to (\xi, \beta, \iota)$ in probability, and completes the proof. □

3 Hard EDF networks

In §3.1 the fluid model equations and the queueing model are presented, as well as the main results, asserting uniqueness and convergence. In §3.2 auxiliary lemmas are established. The proofs of the main results are presented in §3.3.

3.1 Model and main results

3.1.1 Fluid model

The hard version of the policy differs from the soft version in that, when the deadline expires, fluid in the queue reneges. Therefore, we distinguish between the reneging process and the departure process:

$\beta^{is} \in \mathbb{D}_{\mathcal{M}}^{\uparrow}$: $\beta_t^{is}[0,x]$ is the mass with deadlines in $[0,x]$ that has left the i-th queue by time t due to service,

$\beta^{ir} \in \mathbb{D}_{\mathcal{M}}^{\uparrow}$: $\beta_t^{ir}[0,x]$ is the mass with deadlines in $[0,x]$ that has left the i-th queue by time t due to reneging.

The deadline is not fixed but is postponed by a constant $\epsilon > 0$ whenever mass migrates to a new service station. We introduce $\gamma = \left(\gamma^i\right)$ defined by

$$\gamma_t^i[0,x] = \beta_t^{i,s}[0,x-\epsilon], \tag{25}$$

(recalling the convention $[a,b]=\emptyset$ if $a>b$), which expresses the mass that has left server i by time t with its updated deadline in $[0,x]$. The processes α, μ, ξ, ι are defined as before.

We now describe the equations governing the fluid model. The fluid that is served and the fluid that reneges from the system sum up to the total fluid that leaves the queue, and the idle effort is the difference between the potential effort and the processed mass, as described by

$$\beta_t^i[0,x]=\beta_t^{i,s}[0,x]+\beta_t^{i,r}[0,x], \tag{26}$$

$$\iota^i(t)=\mu^i(t)-\beta_t^{i,s}[0,\infty). \tag{27}$$

Reneging does not occur prior to deadline, as expressed by

$$\beta_t^r(t,\infty)=0. \tag{28}$$

The content of the i-th queue is given by

$$\xi_t^i[0,x]=\alpha_t^i[0,x]+\sum_{j=1}^K P_{ji}\gamma_t^{j,s}[0,x]-\beta_t^i[0,x]. \tag{29}$$

The work conservation and EDF conditions are expressed through

$$\int \xi_t^i[0,x]\,d\iota^i(t)=0, \qquad 1\le i\le K, \tag{30}$$

$$\int \xi_t^i[0,x]\,d\beta_t^i(x,\infty)=0, \quad 1\le i\le K. \tag{31}$$

To complete the description we define the processes

$$\rho^i(t)=\beta_t^{i,r}[0,t]=\beta_t^{i,r}[0,\infty), \tag{32}$$

$$\sigma^i(t)=\inf\operatorname{supp}\left[\xi_t^i\right]. \tag{33}$$

Note that $\beta_t^{ir}[0,x]=\rho^i(t\wedge x)$. The quantity $\rho^i(t)$ expresses the cumulative reneging from buffer i. The significance of $\sigma^i(t)$, the left edge of the support of ξ^i, is that reneging from buffer i occurs only at times t when $\sigma^i(t)=t$. Thus

$$\xi_t^i[0,t)=0, \qquad 1\le i\le K, \tag{34}$$

$$\int \mathbb{1}_{\{\sigma^i(t)>t\}}d\rho^i(t)=0, \quad 1\le i\le K. \tag{35}$$

We refer to (25)–(35) as the *fluid model equations*. A solution to these equations for data (α,μ) is a tuple $(\xi,\beta,\beta^s,\beta^r,\rho,\iota)$ for which these equations hold. Note that it is possible to recover β and ρ from β^s and β^r, and vice versa. Therefore, we sometimes refer to a tuple (ξ,β,ρ,ι) or $(\xi,\beta^s,\beta^r,\iota)$ as a solution.

Assumption 1. *1. α takes the form $\alpha_t(B)=\xi_{0-}(B)+\hat\alpha_t(B)$, where $\xi_{0-}\in\mathcal{M}_\sim^K$, and $\hat\alpha\in\mathbb{C}_{\mathcal{M}_\sim}^{\uparrow K}$ satisfies for every $1\le i\le K$*

$$\hat\alpha_t^i(B)=\int_0^t a_s^i(B)ds, \quad t\ge 0, B\in\mathcal{B}(\mathbb{R}_+),$$

where for each t and i, a_t^i is a finite measure on $[t,\infty)$, and $t\mapsto a_t^i(B)$ is measurable. Moreover, $\lim_{\delta\downarrow 0}\sup_{s\in[0,t]}a_s^i[s,s+\delta]=0$.

2. μ takes the form

$$\mu(t) = \int_0^t m(s)\, ds, \quad t \geq 0,$$

for a Borel measurable function $m : [0,\infty) \to \mathbb{R}_+^K$ satisfying $\min_i \inf_{s\in[0,t]} m^i(s) > 0$ for every t.

Theorem 3. *Suppose (α, μ) satisfies Assumption 1. Then there exists a unique solution $(\xi, \beta, \beta^s, \beta^r, \rho, \iota)$ in $\mathbb{D}_{\mathcal{M}}^K \times (\mathbb{D}_{\mathcal{M}}^{K\uparrow})^3 \times (\mathbb{D}_+^{K\uparrow})^2$ to the fluid model. Moreover, β^s satisfies Assumption 1.1.*

We prove this result in §3.3.

3.1.2 Queueing model, scaling

There are many ingredients of the model that are the same as in §2, which we will not repeat. The processes $\alpha_t^{i,N}$, $\mu^{i,N}(t)$, $\xi_t^{i,N}$, $\iota^{i,N}(t)$ have the same meaning as in the soft EDF model. $\beta_t^{i,s,N}[0,x]$ is the number of jobs that have left the i-th queue by time t with deadlines in $[0,x]$ to get served, $\beta_t^{i,r,N}[0,x]$ is the number of jobs that have reneged from the i-th queue by time t with deadlines in $[0,x]$ when their deadlines had expired. Also, the cumulative reneging count is denoted by $\rho^{i,N}(t)$ and satisfies $\rho^{i,N}(t) = \beta_t^{i,r,N}[0,\infty)$.

As already mentioned, the deadline of a job is postponed by ε every time it migrates from one station to another. It is possible that the deadline of a job at a station expires while it is in service at that station. In this case, it is assumed that the job does not renege (in agreement with the model from [4]). A more complicated scenario is when the work associated with this job is sufficiently large that it is not complete even ε units of time after this deadline expires. At this time the deadline at the next station also expires. We prefer not to deal with this possibility because it complicates the notation. Because ε is fixed whereas the service times are downscaled, an event like that becomes less and less probable as N increases. Our assumption, that will allow us to avoid this scenario altogether, is that the service time distributions have bounded supports. This assures that for N large enough this scenario does not occur. The description of the model equations is given for sufficiently large N.

Next, the processes $S^i(t)$, $D^{i,N}(t)$, $\gamma_t^{ij,N}$, $B_t^{i,N}$ and $T^{i,N}(t)$ as well as the relations between them are as in the soft EDF model. Note that the assumption regarding service time distribution supports is really an assumption about the processes S^i.

We turn to mathematically describe the relations between the processes. We have

$$T^{i,N}(t) = \int_0^t B_s^{i,N}[0,\infty)\, d\mu^{i,N}(s),$$
$$\iota^{i,N}(t) = \mu^{i,N}(t) - T^{i,N}(t).$$

$$\beta_t^{i,N}[0,x] = \beta_t^{is,N}[0,x] + \beta_t^{ir,N}[0,x], \tag{36}$$
$$D^{i,N}(t) = \gamma_t^{i,N}[0,\infty) = S^i\left(T^{i,N}(t)\right) - 1, \tag{37}$$
$$\gamma_t^{i,N}[0,x] = \beta_t^{i,s,N}[0,x-\epsilon] - B_t^{i,N}[0,x-\epsilon] + B_{0-}^{i,N}[0,x-\epsilon]. \tag{38}$$

$$\xi_t^{i,N}[0,x] = \alpha_t^{i,N}[0,x] + \sum_{j=1}^K \gamma_t^{ji,N}[0,x] - \beta_t^{i,N}[0,x]. \tag{39}$$

The work conservation condition, the EDF policy, the hard EDF condition and the latest reneging condition are expressed through

$$\int \xi_t^{i,N}[0,x]\, d\iota^{i,N}(t) = 0, \qquad 1 \le i \le K \tag{40}$$

$$\int \xi_t^{i,N}[0,x]\, d\beta_t^{i,N}(x,\infty) = 0, \quad 1 \le i \le K, \tag{41}$$

$$\xi_t^{i,N}[0,t] = 0, \qquad 1 \le i \le K, \tag{42}$$

$$\int \mathbb{1}_{\{\sigma^{i,N}(t)>t\}} d\rho^{i,N}(t) = 0, \qquad 1 \le i \le K, \tag{43}$$

where $\sigma^{i,N}(t) = \inf \operatorname{supp} \xi_t^{i,N}$. Finally, the assumptions on $\pi^{i,N}(n)$, the definition of $\theta^{ij,N}(n)$, $\tau_n^{i,N}$, $\hat{\theta}^{ij,N}(t)$ and the relations between these processes are as in the soft EDF model.

Let us define the error processes as

$$\begin{aligned} e^{i,N}(t) &= \beta_t^{i,s,N}[0,\infty) + \iota^{i,N}(t) - \mu^{i,N}(t), \quad 1 \le i \le K, \\ E^{ij,N}(t,x) &= \gamma_t^{ij,N}[0,x] - P_{ij}\gamma_t^{i,N}[0,x], \quad 1 \le i,j \le K, x \in \mathbb{R}_+. \end{aligned} \tag{44}$$

Theorem 4. *Assume $(\bar{\alpha}^N, \bar{\mu}^N) \Rightarrow (\alpha, \mu)$, where (α, μ) satisfies Assumption 1. Then*

$$\left(\bar{\xi}^N, \bar{\beta}^N, \bar{\beta}^{s,N}, \bar{\beta}^{r,N}, \bar{\iota}^N, \bar{\rho}^N, \bar{e}^N, \bar{E}^N\right) \Rightarrow (\xi, \beta, \beta^s, \beta^r, \iota, \rho, 0, 0), \tag{45}$$

where $(\xi, \beta, \beta^s, \beta^r, \iota, \rho)$ is the unique solution of the fluid model equations corresponding to (α, μ).

3.2 Auxiliary lemmas

When setting $K = 1$ and $P = 0$, the model coincides with the single server model as described in [4], for which uniqueness of the solution of the fluid model equation and convergence have been established in [4].

Lemma 2. *Let $K = 1$ and $P = 0$. Suppose (α, μ) satisfies Assumption 1. Then there exists a unique solution $(\xi, \beta, \iota, \rho)$ to the fluid model equations, and the solution lies in $\mathbb{C}_{\mathcal{M}_\sim} \times \mathbb{C}^{\uparrow}_{\mathcal{M}_\sim} \times \mathbb{C}^{\uparrow} \times \mathbb{C}^{\uparrow}$. Assume, moreover, that $(\bar{\alpha}^N, \bar{\mu}^N) \Rightarrow (\alpha, \mu)$. Then $\left(\bar{\xi}^N, \bar{\beta}^N, \bar{\iota}^N, \bar{\rho}^N\right) \Rightarrow (\xi, \beta, \iota, \rho)$.*

Proof. This is the content of Theorems 4.10 and 5.4 in [4]. □

Lemma 3. *Suppose (α, μ) satisfies Assumption 1 and fix $\tau > 0$. Let $\alpha^\circ \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$ be defined by $\alpha_t^\circ(A) = \alpha_t(A \cap [0,\tau])$ for all $A \in \mathcal{B}(\mathbb{R}_+)$ and $t \in \mathbb{R}_+$. Let $(\xi^\circ, \beta^\circ, \iota^\circ, \rho^\circ)$ and $(\xi, \beta, \iota, \rho)$ be the unique solutions of the fluid model equations (for $K = 1$) corresponding to (α°, μ) and (α, μ), respectively. Then for all $x \in [0,\tau]$, $t \in [0,\tau]$ one has $\rho(t) = \rho^\circ(t)$, $\beta_t^r[0,x] = \beta_t^{\circ,r}[0,x]$, $\beta_t^s[0,x] = \beta_t^{\circ,s}[0,x]$, and $\xi_t[0,x] = \xi_t^\circ[0,x]$.*

Proof. Recall from the proof of Theorem 1 the notation $\Gamma = (\Gamma_1, \Gamma_2)$ for the SM in the finite dimensional orthant. We denote by $\Gamma^{(1)} = (\Gamma_1^{(1)}, \Gamma_2^{(1)})$ the special case where the dimension is 1. That is, $\Gamma^{(1)}$ is merely the SM on the half line. The proof uses crucially the non-anticipation property of this SM [11, p.165], stating that if $\varphi_i = \Gamma^{(1)}(\psi_i)$ for $i = 1, 2$, and for some $T > 0$, $\psi_1 = \psi_2$ on $[0,T]$, then also $\varphi_1 = \varphi_2$ on $[0,T]$.

Next, recall the notation Θ for the VMVSM. Once again, in the special case $K = 1$ we denote this map by $\Theta^{(1)}$. In this proof, t is always assumed to be in $[0,\tau]$. From Lemma 2, using equations (26)-(33) to

reach conditions 1-4 in Definition 1 with the data $(\alpha, \mu+\rho)$, it follows that $(\xi^\circ, \beta^\circ, \iota^\circ, \rho^\circ)$ and $(\xi, \beta, \iota, \rho)$ are respectively the unique solutions of the following two problems (P0) and (P1):

$$\begin{cases} (i) & (\xi^\circ, \beta^\circ, \iota^\circ) = \Theta^{(1)}(\alpha^\circ, \mu+\rho^\circ), \\ (ii) & \xi_t^\circ[0,t) = 0, \\ (iii) & \int \mathbb{1}_{\{\sigma^\circ(t)>t\}} d\rho^\circ(t) = 0, \quad \sigma^\circ(t) = \inf \operatorname{supp}[\xi_t^\circ]. \end{cases} \tag{P0}$$

$$\begin{cases} (i) & (\xi, \beta, \iota) = \Theta^{(1)}(\alpha, \mu+\rho), \\ (ii) & \xi_t[0,t) = 0, \\ (iii) & \int \mathbb{1}_{\{\sigma(t)>t\}} d\rho(t) = 0, \quad \sigma(t) = \inf \operatorname{supp}[\xi_t]. \end{cases} \tag{P1}$$

Consider now the data $(\alpha, \mu+\rho^\circ) \in \mathbb{D}^\uparrow_{\mathcal{M}} \times \mathbb{D}^\uparrow$, and denote the associated unique solution $\left(\hat{\xi}, \hat{\beta}, \hat{\iota}\right) = \Theta^{(1)}(\alpha, \mu+\rho^\circ)$, and $\hat{\sigma}(t) = \inf \operatorname{supp}\left[\hat{\xi}_t\right]$. We will show that the tuple $\left(\hat{\xi}, \hat{\beta}, \hat{\iota}, \rho^\circ\right)$ satisfies $(i)-(iii)$ in (P1). Uniqueness implies then $(\xi, \beta, \iota, \rho) = \left(\hat{\xi}, \hat{\beta}, \hat{\iota}, \rho^\circ\right)$.

$\left(\hat{\xi}, \hat{\beta}, \hat{\iota}, \rho^\circ\right)$ satisfies condition (P1.i) by the definition of $\left(\hat{\xi}, \hat{\beta}, \hat{\iota}\right)$.

To show that $\left(\hat{\xi}, \hat{\beta}, \hat{\iota}, \rho^\circ\right)$ satisfies (P1.ii), use [4, Lemma 2.7] for $x \leq \tau$:

$$\hat{\xi}[0,x] = \Gamma_1^{(1)}(\alpha[0,x] - \mu - \rho^\circ) = \Gamma_1^{(1)}(\alpha^\circ[0,x] - \mu - \rho^\circ) = \xi^\circ[0,x].$$

This implies immediately $\hat{\xi}_t[0,t) = \xi_t^\circ[0,t) = 0$ by (P0.ii).

We need to show now that $\int \mathbb{I}_{\{\hat{\sigma}(t)>t\}} d\rho^\circ(t) = 0$. This will follow from (P0.iii) once we show that

$$\{t \leq \tau : \quad \sigma^\circ(t) = t\} = \{t \leq \tau : \quad \hat{\sigma}(t) = t\}.$$

We already showed that for x and t in $[0,\tau]$: $\hat{\xi}_t[0,x] = \xi_t^\circ[0,x]$. Consider any $t' \in \{t \leq \tau : \sigma^\circ(t) = t\}$. Then $t' = \inf \operatorname{supp}\left[\xi_{t'}^\circ\right] = \inf \operatorname{supp}\left[\hat{\xi}_{t'}\right]$. So t' is in $\{t \leq \tau : \hat{\sigma}(t) = t\}$ and therefore

$$\{t \leq \tau : \quad \sigma^\circ(t) = t\} \subseteq \{t \leq \tau : \quad \hat{\sigma}(t) = t\}.$$

The other direction is proved similarly.

To conclude, when we consider only times in the interval $[0,\tau]$, the tuple $\left(\hat{\xi}, \hat{\beta}, \hat{\iota}, \rho^\circ\right)$ satisfies all the conditions in (P1). So, by uniqueness, $\rho(t) = \rho^\circ(t)$ for $t \leq \tau$.

This implies all the other equalities: $\beta_t^r[0,x] = \rho(t \wedge x) = \rho^\circ(t \wedge x) = \beta_t^\circ[0,x]$ for $x, t \in [0,\tau]$.

Also, for $x, t \in [0,\tau]$, by the non-anticipation property,

$$\xi_t[0,x] = \Gamma_1^{(1)}(\alpha[0,x] - \mu - \rho)(t) = \Gamma_1^{(1)}(\alpha^\circ[0,x] - \mu - \rho^\circ)(t) = \xi_t^\circ[0,x].$$

Finally, for $x, t \in [0,\tau]$: $\beta_t[0,x] = \alpha_t[0,x] - \xi_t[0,x] = \alpha_t^\circ[0,x] - \xi_t^\circ[0,x] = \beta_t^\circ[0,x]$, and $\beta_t^s[0,x] = \beta_t[0,x] - \beta_t^r[0,x] = \beta_t^\circ[0,x] - \beta_t^{\circ,r}[0,x] = \beta_t^{\circ,s}[0,x]$. □

The next lemma states that the fluid model equations for $K = 1$ 'preserve' Assumption 1.

Lemma 4. *Let $K = 1$ and $P = 0$ and suppose (α, μ) satisfies Assumption 1. Let $(\xi, \beta, \iota, \rho)$ be the unique solution of the fluid model equations. Let γ_t be defined via the relation $\gamma_t[0, x] = \beta_t^s[0, x - \epsilon]$. Then γ also satisfies Assumption 1.1 with zero initial condition, i.e. $\gamma \in \mathbb{C}^{\uparrow}_{\mathcal{M}_\sim}$*

$$\gamma_t(B) = \int_0^t g_s(B)ds, \quad t \geq 0, B \in \mathcal{B}(\mathbb{R}_+), \tag{46}$$

where for each t, g_t is a finite measure on $\mathbb{R}_+$ with $g_t[0, t) = 0$, and the mapping $t \mapsto g_t(B)$ is measurable. Moreover, $\lim_{\delta \downarrow 0} \sup_{s \in [0,t]} g_s[s, s + \delta] = 0$.

Proof. The proof is based mainly on the following two facts

$$\beta_t^s(B) \leq \alpha_t(B), \qquad B \in \mathcal{B}(\mathbb{R}), \tag{47}$$

$$\beta_t^s(\mathbb{R}_+) - \beta_\tau^s(\mathbb{R}_+) \leq \int_\tau^t m(s)ds, \qquad 0 \leq \tau < t, \tag{48}$$

and uses disintegration.

By the fluid model equations for $K = 1$, using (29) (recalling $P = 0$) and (26), for a Borel set B we have $\alpha_t(B) = \xi_t(B) + \beta_t(B)$ as well as $\beta_t(B) = \beta_t^s(B) + \beta_t^r(B)$, proving (47). Next, by (27), the monotonicity of ι and Assumption 1.2, (48) holds.

We first show that γ belongs to $\mathbb{C}^{\uparrow}_{\mathcal{M}_\sim}$. From the axioms of our model $\beta^s \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$. By (47) and the assumption $\alpha_t \in \mathcal{M}_\sim$ it follows that $\beta_t^s \in \mathcal{M}_\sim$ for every t. The continuity $t \mapsto \beta_t^s$ follows from (48). This shows that β^s, and in turn, γ, belongs to $\mathbb{C}^{\uparrow}_{\mathcal{M}_\sim}$.

Consider now the space $\mathbb{R}_+^2$ with its Borel σ-algebra. Let λ be the measure on this space determined by γ via the relation

$$\lambda([x, y] \times [s, t]) = \gamma_t[x, y] - \gamma_s[x, y], \qquad x < y, \ s < t.$$

By (48), the measure $\lambda(\mathbb{R}_+ \times dt)$ is dominated by the measure $m(t)dt$. Let $\mathbb{T} : \left(\mathbb{R}_+^2, \mathcal{B}(\mathbb{R}_+^2)\right) \to (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ be defined by $\mathbb{T}(x, t) = t$. We use the disintegration theorem [10, Theorem 1], with $\mathbb{T}$ as the measurable map. According to this theorem there exists a family of finite measures $\tilde{g}_t$ on $\mathbb{R}_+$ such that $t \mapsto \tilde{g}_t(B)$ is measurable for every $B \in \mathcal{B}(\mathbb{R}_+)$, and for each nonnegative measurable f,

$$\int f(x, t)\lambda(dx, dt) = \int f(x, t)\tilde{g}_t(dx)m(t)dt. \tag{49}$$

Denote $\hat{g}_t(B) = \tilde{g}_t(B)m(t)$. It will be shown that

$$\hat{g}_t[0, t + \varepsilon) = 0 \quad \text{ for a.e. } t. \tag{50}$$

Once the above is established, we can set $g_t(B) = \hat{g}_t(B \cap [t + \varepsilon, \infty))$, $B \in \mathcal{B}(\mathbb{R}_+)$, by which we achieve for a.e. t, $g_t(B) = \hat{g}_t(B)$. Hence in view of (49) one has (46). As for the two final assertions made in the lemma, we certainly have $g_t[0, t) \leq g_t[0, t + \varepsilon) = 0$, and for all small δ, $\sup_s g_s[s, s + \delta] = 0$, by which these assertions are true.

We thus turn to showing (50). Fix $x \leq t \leq t_0$. From the hard EDF equations and the assumption that jobs arrive before their deadlines:

$$\begin{aligned}
&0 = \xi_t[0, x] = \alpha_t[0, x] - \beta_t[0, x], \\
&\alpha_{t_0}[0, x] - \alpha_t[0, x] = 0, \\
&\beta_{t_0}^r[0, x] - \beta_t^r[0, x] = \rho(x \wedge t_0) - \rho(x \wedge t) = \rho(x) - \rho(x) = 0,
\end{aligned} \tag{51}$$

$$\Rightarrow \beta^s_{t_0}[0,x] - \beta^s_t[0,x] = \alpha_{t_0}[0,x] - \alpha_t[0,x] - \beta^r_{t_0}[0,x] + \beta^r_t[0,x] = 0.$$

This, together with (25) that expresses the fact that deadlines are postponed by ε after service, implies that every rectangular subset $[0,t+\varepsilon]\times[t,t_0]$ of the set $\{(x,t)\in\mathbb{R}^2_+ : x < t+\varepsilon\}$ satisfies:

$$\begin{aligned}\lambda\left([0,t+\varepsilon]\times[t,t_0]\right) &= \gamma_{t_0}[0,t+\varepsilon] - \gamma_t[0,t+\varepsilon] \\ &= \beta^s_{t_0}[0,t] - \beta^s_t[0,t] \\ &= 0.\end{aligned} \tag{52}$$

And, since $\{(x,t)\in\mathbb{R}^2_+ : x < t+\epsilon\}$ is contained in a countable union of rectangles of this form,

$$0 = \lambda\left(\{(x,t)\in\mathbb{R}^2_+ : x < t+\varepsilon\}\right) = \int_0^\infty \hat{g}_s[0,s+\varepsilon)\,ds,$$

and (50) follows. □

Remark. Note that for any finite collection $\{\gamma^i\}_{i=1}^K$ and a set of positive coefficients $\{P_i\}_{i=1}^K$, if each element of the collection satisfies Assumption 1, then does also $\sum_{i=1}^K P_i\gamma^i$. Lemma 4 considers a single server, but we will use the conclusion on the sum when we will show that the total arrival process, the sum of exogenous and the endogenous arrival processes, satisfies the assumptions.

3.3 Proof of main results

This section opens with the proof of Theorem 3 regarding existence and uniqueness of solutions to the fluid model equations.

Proof of Theorem 3. Denote

$$\begin{aligned}\Xi &= (\xi,\beta,\beta^s,\beta^r,\gamma,\iota,\rho), \\ \mathbb{X} &= \mathbb{D}^K_{\mathcal{M}}\times\mathbb{D}^{\uparrow K}_{\mathcal{M}}\times\mathbb{D}^{\uparrow K}_{\mathcal{M}}\times\mathbb{D}^{\uparrow K}_{\mathcal{M}}\times\mathbb{D}^{\uparrow K}_{\mathcal{M}}\times\mathbb{D}^{\uparrow K}\times\mathbb{D}^{\uparrow K},\end{aligned}$$

$$\mathcal{X} = \{\Xi\in\mathbb{X}: \quad \text{The components of } \Xi \text{ satisfy (25)-(28),(30)-(35)}\}.$$

Recall that the fluid model equations are (25)-(35). Equation (29), the only one excluded from the definition of $\mathcal{X}$, is the equation that couples between the different servers. Indeed, the remaining equations are fluid model equations for K separate single server systems. This exclusion makes it convenient to use the single server results in our proof. The existence proof, provided first, will be complete once we construct a tuple $\Xi\in\mathcal{X}$ that satisfies (29). It is then shown that this tuple is unique. Both existence and uniqueness are proved by induction.

First consider, for each i, the fluid model solution of a single server with primitives $\left(\alpha^i(\cdot\cap[0,\varepsilon)),\mu^i\right)$, denoted $\left(\xi^{i,(1)},\beta^{i,(1)},\iota^{i,(1)},\rho^{i,(1)}\right)$. Denote also $\gamma^{i,(1)}_t[0,x] = \beta^{is,(1)}_t[0,x-\varepsilon]$. Then, for $n>1$, denote the fluid model solution of a single server with primitives $\left(\alpha^i(\cdot\cap[0,n\varepsilon)) + \sum_{j=1}^K P_{ji}\gamma^{j,(n-1)}(\cdot\cap[0,n\varepsilon)),\mu^i\right)$ by $\left(\xi^{i,(n)},\beta^{i,(n)},\iota^{i,(n)},\rho^{i,(n)}\right)$ and $\gamma^{i,(n)}_t[0,x] = \beta^{i,(n)}_t[0,x-\varepsilon]$. This inductively defines the tuples $\Xi^{(n)}$. Note that these tuples belong to $\mathcal{X}$ as solutions to the fluid model of hard EDF servers. Note also that all $\beta^{is,(n)}$ satisfy Assumption 1. We now show that these tuples are consistent with each other

in that for $(x,t) \in [0,n\varepsilon)^2$ and $m > n$, we have $\beta_t^{is,(n)}[0,x] = \beta_t^{is,(m)}[0,x]$, $\rho^{i,(n)}(t) = \rho^{i,(m)}(t)$ and $\xi_t^{i,(n)}[0,x] = \xi_t^{i,(m)}[0,x]$.

It is proved by induction over n that these identities hold for all $m > n$. If $(x,t) \in [0,\varepsilon)^2$, then $\gamma_t^{j,(m)}([0,x] \cap [0,m\varepsilon)) = \gamma_t^{j,(m)}[0,x] = 0$ and $\alpha_t^i([0,x] \cap [0,m\varepsilon)) = \alpha_t^i([0,x] \cap [0,\varepsilon))$, so the data generating $\rho^{(1)}$, $\rho^{(m)}$, $\xi^{(1)}$, $\xi^{(m)}$, $\beta^{s,(1)}$ and $\beta^{s,(m)}$ coincide on $[0,\varepsilon)^2$ and Lemma 3 yields the desired conclusion for $n = 1$. Assuming the claim is true for n, consider $m \geq n+1$ and $(x,t) \in [0,(n+1)\varepsilon)^2$. Then

$$\alpha_t^i([0,x] \cap [0,(n+1)\varepsilon)) = \alpha_t^i([0,x] \cap [0,m\varepsilon)) = \alpha_t^i[0,x].$$

If $t < n\varepsilon$, then the induction assumption implies

$$\gamma_t^{j,(n)}([0,x] \cap [0,(n+1)\varepsilon)) = \beta_t^{js,(n)}[0,x-\varepsilon] = \beta_t^{js,(m)}[0,x-\varepsilon] = \gamma_t^{j,(m)}([0,x] \cap [0,(m+1)\varepsilon)).$$

If $t \in [n\varepsilon,(n+1)\varepsilon)$, by (51), we have that whenever $t > x$: $\beta_t^{i,s}[0,x] = \beta_x^{i,s}[0,x]$. It follows that if $t \in [n\epsilon,(n+1)\epsilon)$,

$$\begin{aligned}\gamma_t^{j,(n)}([0,x] \cap [0,(n+1)\varepsilon)) &= \beta_t^{js,(n)}[0,x-\epsilon] = \beta_{[x-\epsilon]^+}^{js,(n)}[0,x-\epsilon] \\ &= \beta_{[x-\epsilon]^+}^{js,(m)}[0,x-\epsilon] = \gamma_t^{j,(m)}[0,x] = \gamma_t^{j,(m)}([0,x] \cap [0,(m+1)\varepsilon)).\end{aligned}$$

So, the data generating $\rho^{(n)}$, $\rho^{(m)}$, $\xi^{(n)}$, $\xi^{(m)}$, $\beta^{s,(n)}$ and $\beta^{s,(m)}$ coincide on $[0,(n+1)\varepsilon)^2$ and Lemma 3 yields the desired conclusion for $n+1$.

The above shows that, for example, $\beta_t^{i,(n)}[0,x]$ becomes constant as n increases. Hence we may extract from the sequence a tuple in the following way. Given x and t, let $n > (x \vee t)/\varepsilon$ and let $\beta_t^i[0,x] = \beta_t^{i,(n)}[0,x]$, and $\xi_t^i[0,x] = \xi_t^{i,(n)}[0,x]$. For every fixed t, this uniquely defines measures by determining their values on all sets $[0,x]$, a collection generating the σ-algebra $\mathcal{B}(\mathbb{R}_+)$. Similarly, for each t, $\rho_t^{(n)}$ becomes constant as n gets large, hence we let $\rho = \lim_n \rho^{(n)}$. We also define $\beta_t^{i,r}[0,x] = \rho^i(x \wedge t)$, $\beta^{i,s} = \beta^i - \beta^{i,r}$, $\sigma^i(t) = \inf \operatorname{supp}\left[\xi_t^i\right]$, $\iota^i = \mu^i - \beta^{i,s}[0,\infty)$ and $\gamma_t^i[0,x] = \beta_t^{i,s}[0,x-\varepsilon]$.

The tuple (ξ,β,ι,ρ) thus constructed is our candidate, and we now show that it is indeed a solution to the fluid model equations.

First, by construction, Ξ satisfies (29). (25), (26), (27), (32) and (33) hold by definition, and (34) is immediate by construction. (28) is also immediate by $\beta_t^{i,r}(t,\infty) = \rho^i(t) - \rho^i(t \wedge t) = 0$. Equations (30) and (31) are satisfied because, as will soon be shown, $\psi_x^{i,(n)} := \beta^{is,(n)}(x,\infty) + \iota^{i,(n)} - \beta^{is}(x,\infty) - \iota^i \in \mathbb{D}^\uparrow$ and $\xi_t^{i,(n)}[0,x] = \xi_t^i[0,x]$ for large enough n. Recall also that $\Xi^{(n)} \in \mathcal{X}$ and (40) and (41) . Therefore,

$$\begin{aligned}&\int \xi_s^i[0,x]\, d\beta_s^{i,s}(x,\infty) + \int \xi_s^i[0,x]\, d\iota^i(s) \\ &= \lim_{n\to\infty} \left| \int \xi_s^{i,(n)}[0,x]\, d\beta_s^{is,(n)}(x,\infty) + \int \xi_s^{i,(n)}[0,x]\, d\iota^{i,(n)}(s) - \int \xi_s^i[0,x]\, d\beta_s^{is}(x,\infty) - \int \xi_s^i[0,x]\, d\iota^i(s) \right| \\ &= \lim_{n\to\infty} \left| \int \xi_s^i[0,x]\, d\psi_x^{i,(n)}(s) \right| \\ &\leq \left\|\xi^i[0,x]\right\|_T \lim_{n\to\infty} \left(\beta_T^{is,(n)}(x,\infty) + \iota^{i,(n)}(T) - \beta_T^{is}(x,\infty) - \iota^i(T)\right) = 0. \end{aligned} \tag{53}$$

For the equality in (53), note that $\beta_T^{i,s}[0,\infty) = \lim_m \lim_n \beta_T^{is,(n)}[0,m]$ while $\beta_T^{is,(n)}[0,m]$ is monotone in both n and m, hence interchanging the limits is justified. Thus we have $\beta_T^{is,(n)}[0,\infty) \to \beta_T^{i,s}[0,\infty)$ and $\iota^{i,(n)}(T) \to \iota^i(T)$ and (53).

Lemma 2.2 in [4] implies $\psi_x^{i,(n)} \in \mathbb{D}^{\uparrow}$ and the monotonicity in n of $\beta_t^{is,(n)}[0,x]$, provided one shows that for all n

$$\alpha^i[x\wedge n\varepsilon, x\wedge(n+1)\varepsilon) + \sum_{j=1}^K P_{ji}\left(\gamma^{j,(n)}[0,x\wedge(n+1)\varepsilon) - \gamma^{j,(n-1)}[0,x\wedge n\varepsilon)\right) \in \mathbb{D}^{\uparrow}.$$

The first term is in $\mathbb{D}^{\uparrow}$ by assumption, and the second term is now shown to belong to $\mathbb{D}^{\uparrow}$ by induction, where each step invokes Lemma 2.2 from [4]. Indeed, the lemma yields

$$\beta^{is,(2)}[0,x\wedge 2\varepsilon) - \beta^{is,(1)}[0,x\wedge\varepsilon) = \mu^i - \iota^{i,(2)} - \beta^{is,(2)}[x\wedge 2\varepsilon,\infty) - \mu^i + \iota^{i,(1)} + \beta^{is,(1)}[x\wedge\varepsilon,\infty) \in \mathbb{D}^{\uparrow}$$

because $\alpha^i[x\wedge\varepsilon, x\wedge 2\varepsilon) + \sum_{j=1}^K P_{ji}\gamma^{j,(1)}[0,x\wedge 2\varepsilon) \in \mathbb{D}^{\uparrow}$. Also,

$$\begin{aligned}&\beta^{is,(n)}[0,x\wedge n\varepsilon) - \beta^{is,(n-1)}[0,x\wedge(n-1)\varepsilon)\\ &\qquad = \mu^i - \iota^{i,(n)} - \beta^{is,(n)}[x\wedge n\varepsilon,\infty) - \mu^i + \iota^{i,(n-1)} + \beta^{is,(n-1)}[x\wedge(n-1)\varepsilon,\infty) \in \mathbb{D}^{\uparrow}\end{aligned}$$

if $\alpha^i[x\wedge(n-1)\varepsilon, x\wedge n\varepsilon) + \sum_{j=1}^K P_{ji}\left(\gamma^{j,(n-1)}[0,x\wedge n\varepsilon) - \gamma^{j,(n-2)}[0,x\wedge(n-1)\varepsilon)\right) \in \mathbb{D}^{\uparrow}$.

For (35), note that

$$\{t:\sigma^i(t)>t\} \subset \bigcup_{n\geq m}\{t:\sigma^{i,(n)}>t\} \quad \forall m.$$

This is due to the fact that if $\xi_t^i[0,x]>0$ then also $\xi_t^{i,(n)}[0,x]>0$ for all $n>(x\vee t)/\varepsilon$. Consider $t\in[0,T]$ for T fixed. Recall that $\rho_t^i = \rho_t^{i,(n)}$ for all $t\in[0,T]$, $n>T/\varepsilon$. Using the above display and the union bound, we obtain for all large n,

$$\int_{[0,T]} \mathbb{1}_{\{\sigma^i(s)>s\}} d\rho^i(s) \leq \sum_{n>T/\varepsilon}\int \mathbb{1}_{\left\{\sigma^{i,(n)}(s)>s\right\}} d\rho^i(s) = \sum_{n>T/\varepsilon}\int \mathbb{1}_{\left\{\sigma^{i,(n)}(s)>s\right\}} d\rho^{i,(n)}(s) = 0.$$

This shows $\Xi\in\mathcal{X}$ and completes the existence proof.

The argument for uniqueness is as follows. We show that for every $x,t\in\mathbb{R}_+$, the quantities $\xi_t[0,x]$, $\beta_t[0,x]$, $\rho(t)$ and $\iota(t)$ are uniquely determined by the primitives $(\alpha,\mu)\in\mathbb{D}_{\mathcal{M}}^{\uparrow K}\times\mathbb{D}^{\uparrow K}$. This we do by arguing that the claim holds for every $(x,t)\in[0,n\varepsilon)^2$, by induction in n.

First, directly from (29), for $x\in[0,\varepsilon)$, for the i-th server:

$$\xi_t^i[0,x] = \alpha_t^i[0,x] - \beta_t^i[0,x].$$

In addition, $\left(\xi^i,\beta^i,\iota^i,\rho^i\right)\in\mathcal{X}$. By Lemma 3, $\xi_t^i[0,x]$, $\beta_t^{is}[0,x]$ and $\rho^i(t)$ coincide with the unique solution to the fluid model equations of a single server with primitives $(\alpha(\cdot\cap[0,\varepsilon)),\mu)$ on $[0,\varepsilon)^2$. In particular, for $t,x\in[0,\varepsilon)$, $\beta_t^{is}[0,x]$ can be written as required in Assumption 1.1.

Next, assume that the uniqueness statement holds for $(x,t)\in[0,n\varepsilon)^2$. Let $\left(\xi^i,\beta^i,\iota^i,\rho^i\right)$ denote the unique tuple satisfying for all $1\leq i\leq K$ and $(x,t)\in[0,n\varepsilon)^2$,

$$\begin{aligned}&\xi_t^i[0,x] = \alpha_t^i[0,x] + \sum_{j=1}^K P_{ji}\beta_t^{j,s}[0,x-\varepsilon] - \beta_t^i[0,x],\\ &\left(\xi^i,\beta^i,\iota^i,\rho^i\right)\in\mathcal{X}.\end{aligned}$$

Assume in addition that for those x and t, $\beta_t^{is}[0,x]$ can be written as required in Assumption 1.1.

Consider now $(x,t) \in [0,(n+1)\varepsilon)^2$. First, directly from Equation (29), for the i-th server:

$$\xi_t^i[0,x] = \alpha_t^i[0,x] + \sum_{j=1}^K P_{ji}\beta_t^{j,s}[0,x-\varepsilon] - \beta_t^i[0,x],$$
$$\left(\xi^i,\beta^i,\iota^i,\rho^i\right) \in \mathcal{X}.$$

If x is in $[0,(n+1)\varepsilon)$ then $x-\varepsilon < n\varepsilon$. Therefore, $\beta_t^{i,s}[0,x-\varepsilon]$ is uniquely determined by (α,μ) for $t \in [0,n\varepsilon)$ by our induction assumption. Now for $t \in [n\varepsilon,(n+1)\varepsilon)$, by (51), we have that whenever $t > x$: $\beta_t^{i,s}[0,x] = \beta_x^{i,s}[0,x]$. It follows that if $t \in [n\epsilon,(n+1)\epsilon)$ then for $1 \le j \le K$: $\beta_t^{j,s}[0,x-\epsilon] = \beta_{[x-\epsilon]^+}^{j,s}[0,x-\epsilon]$, which is uniquely determined, in view of the induction assumption.

By Lemma 3, $\left(\xi^i,\beta^i,\rho^i\right)$ coincides on $[0,(n+1)\epsilon)^2$ with the unique solution of the fluid model equations for a single server with primitives $\left(\alpha^i(\cdot\cap[0,(n+1)\varepsilon)) + \sum_{j=1}^K P_{ji}\gamma^j(\cdot\cap[0,(n+1)\varepsilon)),\mu\right)$. In particular, $\beta_t^{is}[0,x]$ can be written as required in Assumption 1.1 for x and t in $[0,(n+1)\varepsilon)$ as well.

Finally, ι is uniquely determined for all t via $\iota = \mu+\rho-\beta^s[0,\infty)$. □

Our final goal is to prove convergence to the fluid model equations. We first show that the sequence $\{\bar\rho^N\}$ is tight, and deduce from that tightness of the entire tuple appearing on the LHS of (45). Later we complete the proof by showing that any subsequential limit satisfies the fluid model equations.

Lemma 5. *The sequence $\{\bar\rho^N\}$ is C-tight. Consequently, the tuple $\left(\bar\xi^N,\bar\beta^{s,N},\bar\beta^{r,N},\bar\beta^N,\bar\gamma^N,\bar\iota^N\right)$ is C-tight.*

Proof. The second statement follows from the first by a simple inductive argument over the squares $[0,n\varepsilon)^2$ based on the continuous mapping theorem and Remark 5.1 of [4]. We omit the details.

To prove the first claim, fix T. C-tightness is shown for $\{\bar\rho^N|_{[0,T]}\}$. Define $F_{\bar\alpha^{i,N}}(x) = \bar\alpha_T^{i,N}[0,x]$. To show C-tightness, note first that $\|\bar\rho^N\|_T$ is dominated by $\|\bar\alpha^N\|_T$, that is a tight sequence of RVs. Hence it remains to show that for every $\delta>0$, $w_T(\bar\rho^N,\delta)\to 0$ in probability, as $N\to\infty$. To this end, we bound the aforementioned modulus of continuity in terms of $w_\infty\left(F_{\bar\alpha_T^{i,N}},\delta\right)$ using the following chain of inequalities. The justification of each step in this chain is given below. For $0 \le t \le t+\delta \le T$,

$$\bar\rho^N(t+\delta) - \bar\rho^N(t) \le \bar\alpha_{t+\delta}^N[0,t+\delta] - \bar\alpha_t^N[0,t] + \sum_{j=1}^K \bar\gamma_{t+\delta}^{ji,N}[0,t+\delta] - \sum_{j=1}^K \bar\gamma_t^{ji,N}[0,t] \tag{54}$$

$$\le \bar\alpha_T^{i,N}[t,t+\delta] + \sum_{j=1}^K \bar\gamma_{t+\delta}^{j,N}[t,t+\delta] \tag{55}$$

$$\le \bar\alpha_T^{i,N}[t,t+\delta] + \sum_{j=1}^K \bar\beta_{t+\delta}^{js,N}[t-\varepsilon,t+\delta-\varepsilon] + N^{-1} \tag{56}$$

$$\le \bar\alpha_T^{i,N}[t,t+\delta] + \sum_{n=1}^{\lfloor\frac{t+\delta}{\varepsilon}\rfloor}\sum_{j=1}^K \left(\bar\alpha_T^{j,N}[t-n\varepsilon,t+\delta-n\varepsilon] + \frac{1}{N}\right) + N^{-1} \tag{57}$$

$$\le \left(\frac{KT}{\varepsilon}+1\right)\left(\max_i w\left(F_{\bar\alpha_T^{i,N}},\delta\right) + \frac{1}{N}\right).$$

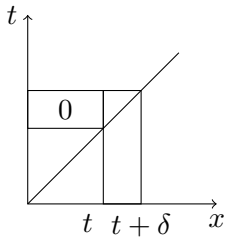

Figure 1: Illustration for Inequality (55).

The convergence $\bar{\alpha}^N \Rightarrow \alpha$ and the continuity of the path $t \mapsto \alpha_t$ we have $\bar{\alpha}_T^N \Rightarrow \alpha_T$. Since by Assumption 1, α_T has no atoms, the continuous, monotone, bounded function $x \mapsto F_{\alpha_T}(x)$ is uniformly continuous, and we have $\lim_{\eta \downarrow 0} \limsup_N P(w_\infty(F_{\bar{\alpha}_T^{i,N}}, \delta) > \eta) = 0$. This shows that $\{\bar{\rho}^N\}$ are C-tight.

It remains to prove the chain of inequalities.

Inequality (54) follows from (39) with $x = t$, $\bar{\xi}_t^{i,N}[0,t] = 0$, and $\bar{\beta}_t^{is,N}[0,t] \leq \bar{\beta}_{t+\delta}^{is,N}[0,t+\delta]$.

For inequality (55), first, $\bar{\alpha}_{t+\delta}^{i,N}[0,t] = \bar{\alpha}_t^{i,N}[0,t]$ because, by assumption, no job enters the system with an overdue deadline (see Figure 1). Hence

$$\begin{aligned}\bar{\alpha}_{t+\delta}^N[0,t+\delta] - \bar{\alpha}_t^N[0,t] &= \bar{\alpha}_{t+\delta}^N[0,t+\delta] - \bar{\alpha}_{t+\delta}^N[0,t] + \bar{\alpha}_{t+\delta}^{i,N}[0,t] - \bar{\alpha}_t^{i,N}[0,t] \\ &\leq \bar{\alpha}_{t+\delta}^N[t,t+\delta] \\ &\leq \bar{\alpha}_T^N[t,t+\delta].\end{aligned}$$

Then, use a similar property for $\bar{\gamma}^{is,N}$: $\bar{\gamma}_{t+\delta}^{is,N}[0,t] = \bar{\gamma}_t^{is,N}[0,t]$. This is due to the fact that no job is served after its deadline has expired. To see this, fix any $x \leq t \leq t_0$, and recall that by (19)

$$\bar{\gamma}_{t_0}^{ji,N}[0,x] - \bar{\gamma}_t^{ji,N}[0,x] = \int_{(t,t_0]} \theta^{ji,N}(s)\, d\bar{\gamma}_s^{i,N}[0,x] = 0. \tag{58}$$

It follows that $\bar{\gamma}_{t+\delta}^{ji,N}[0,t+\delta] - \bar{\gamma}_t^{ji,N}[0,t+\delta] \leq \bar{\gamma}_{t+\delta}^{j,N}[t,t+\delta]$.

Inequality (56) follows from Equation (38).

For (57), we prove by induction that

$$\sum_{j=1}^K \bar{\beta}_T^{js,N}[t,t+\delta] \leq \sum_{n=0}^{\lfloor \frac{t+\delta}{\varepsilon} \rfloor} \sum_{j=1}^K \left(\bar{\alpha}_T^{j,N}[t-n\varepsilon, t+\delta-n\varepsilon] + \frac{1}{N} \right).$$

If t is such that $t + \delta < \varepsilon$, then the statement follows from Equation (39) and $\bar{\beta}_T^{jr,N}[t,t+\delta] \geq 0$.

Now, assume that the desired inequality holds for t such that $\lfloor \frac{t+\delta}{\varepsilon} \rfloor = m - 1$. Then for t such that $\lfloor \frac{t+\delta}{\varepsilon} \rfloor = m$, we have

$$\sum_{j=1}^K \bar{\beta}_T^{js,N}[t-\varepsilon, t+\delta-\varepsilon] \leq \sum_{n=1}^m \sum_{j=1}^K \left(\bar{\alpha}_T^{j,N}[t-n\varepsilon, t+\delta-n\varepsilon] + \frac{1}{N} \right).$$

For these values of t, from (39),

$$\bar{\beta}_T^{is,N}[t,t+\delta] \leq \bar{\alpha}_T^{i,N}[t,t+\delta] + \sum_{j=1}^K \bar{\gamma}_T^{ji,N}[t,t+\delta].$$

Summing over all j, using (38) and the assumption,

$$\begin{aligned}\sum_{i=1}^{K}\bar{\beta}_T^{is,N}[t,t+\delta] &= \sum_{i=1}^{K}\bar{\alpha}_T^{i,N}[t,t+\delta]+\sum_{j=1}^{K}\bar{\gamma}_T^{j,N}[t,t+\delta]\\ &\le \sum_{i=1}^{K}\bar{\alpha}_T^{i,N}[t,t+\delta]+\sum_{j=1}^{K}\bar{\beta}_T^{js,N}[t-\varepsilon,t+\delta-\varepsilon]+\sum_{j=1}^{K}\frac{1}{N}\\ &\le \sum_{i=1}^{K}\bar{\alpha}_T^{i,N}[t,t+\delta]+\sum_{n=1}^{m}\sum_{j=1}^{K}\left(\bar{\alpha}_T^{j,N}[t-n\varepsilon,t+\delta-n\varepsilon]+\frac{1}{N}\right)+\sum_{j=1}^{K}\frac{1}{N}\\ &= \sum_{n=0}^{m}\sum_{j=1}^{K}\left(\bar{\alpha}_T^{j,N}[t-n\varepsilon,t+\delta-n\varepsilon]+\frac{1}{N}\right).\end{aligned}$$

This completes the proof of the chain of inequalities and the result follows. □

This means that the tuple is also sequentially compact. We will use that fact soon to prove Theorem 4.

Proof of Theorem 4. First, the convergence $\left(\bar{e}^N,\bar{E}^N\right)\Rightarrow(0,0)$ follows by the same reasons as for the soft version. Next, from Lemma 5 and Prohorov's theorem, every subsequence of the tuple has a convergent subsequence. We will show that any subsequential limit must be a solution to the fluid model; uniqueness implies then convergence of the entire sequence.

Consider a convergent subsequence and denote by $(\rho,\alpha,\mu,\xi,\beta^s,\beta^r,\beta,\gamma,\iota)$ its limit. We appeal to Skorokhod representation theorem ([7, Thm. 6.7]), and assume without loss of generality convergence a.s. It is now argued that the tuple $(\rho,\alpha,\mu,\xi,\beta^s,\beta^r,\beta,\gamma,\iota)$ satisfies the fluid model equations.

Equation (27) follows by $\bar{e}^{i,N}\Rightarrow 0$. Identities (25), (26), (29), and (32) will follow from the convergence of the tuple once it is shown that for every t, the measures ξ_t, β_t, etc., have no atoms. To show that these measures have no atoms, consider relation (39) with $x<\varepsilon$, in which case the term involving γ^N is absent. Using the assumption that $\alpha\in\mathbb{C}_{\mathcal{M}_\sim}^{\uparrow K}$ and Portmanteau theorem, for every $0\le a<b<\varepsilon$, a.s.,

$$\beta_t^{is}(a,b)\le \liminf\bar{\beta}_t^{is,N}(a,b)\le\liminf\bar{\alpha}_t^{i,N}(a,b)=\alpha_t^i(a,b).$$

Hence the fact that α_t^i has no atoms implies that the same is true for β_t^{is}, on the interval $[0,\varepsilon)$. Hence the same holds for γ_t^i and a similar argument holds for ξ_t^i. An inductive argument over intervals $[0,n\varepsilon)$ used in (39) shows that these measures are all atomless on all of $\mathbb{R}_+$. (The induction argument is omitted).

Moving now to show (30) and (31), note that by (38), (39), (40), (41), and (44),

$$\begin{aligned}&\left(\bar{\xi}^{i,N}[0,x],\bar{\beta}^{is,N}(x,\infty)+\bar{\iota}^{i,N}\right)=\\ &\Gamma^{(1)}\left(\bar{\alpha}^{i,N}[0,x]+\sum_{j=1}^{K}P_{ji}\left(\bar{B}^{i,N}[0,x-\varepsilon]-\bar{B}_{0-}^{i,N}[0,x-\varepsilon]\right)\right.\\ &\quad\left.+\sum_{j=1}^{K}P_{ji}\bar{\beta}^{js,N}[0,x-\varepsilon]-\bar{\rho}^{i,N}(\cdot\wedge x)-\bar{\mu}^{i,N}-\bar{e}^{i,N}+\sum_{j=1}^{K}\bar{E}^{ji,N}(\cdot,x)\right).\end{aligned}$$

Recalling that $\Gamma^{(1)}$ is continuous, using (24) and the fact that there are no atoms, one obtains (30) and (31) by the definition of the Skorokhod map.

It remains to show that the limit satisfies the condition $\int \mathbb{1}_{\{\sigma^i(t)>t\}} d\rho(t) = 0$. The idea is similar to the proof in Section 5.1.4 in [4]; however, there are many details that are different. By Fatou's lemma, it is enough to prove that the event

$$E_0^i = \left\{ \int_0^T \mathbb{1}_{\{\sigma^i(t)>t+\delta\}} d\rho^i(t) > 0 \right\}$$

occurs with probability zero for all $1 \leq i \leq K$. We refer to Lemma 5.9 in [4] and note that there exists a $[0,T) \cup \{\infty\}$-valued random variable τ such that $\mathbb{P}\left(E_0^i\right) = \mathbb{P}\left(E_1^i \cap E_2^i\right)$ where

$$E_1^i = \left\{\tau < T, \sigma^i(\tau) > \tau + \delta\right\}, \qquad E_2^i = \left\{\rho^i(\tau+\delta) > \rho^i(\tau), \forall \delta > 0\right\}.$$

Define

$$E_3^i = \left\{\exists \delta(\omega) > 0 : \rho^i(\tau+\delta) = \rho^i(\tau)\right\}$$

and note that $\mathbb{P}\left(E_1^i \cap E_2^i\right) = \mathbb{P}\left(E_1^i \cap E_3^{ic}\right)$.

We wish to show that for any $\delta > 0$, $\mathbb{P}\left(E_0^i\right) = 0$, which is equivalent to showing $\mathbb{P}\left(E_1^i \cap E_3^{ic}\right) = 0$ for all i. Note that it is enough to show this for any $\varepsilon > \delta > 0$. In fact, we shall take $\delta < \varepsilon \wedge \delta_0$, where $\delta_0 \in (0,1)$ satisfies

$$a_s^i[s, s+2\delta_0] < m^i(s) \text{ for all } s \in [0, T+1] \text{ and } i \in \{1, ..., K\}. \tag{59}$$

The existence of such δ_0 is guaranteed by Assumption 1.

By (39) and (42), for $b > a$,

$$\bar{\rho}^{i,N}(b) - \bar{\rho}^{i,N}(a) + \bar{\beta}_b^{is,N}(a,b] - \bar{\beta}_a^{is,N}(a,b] = \bar{\alpha}_b^{i,N}(a,b] - \bar{\alpha}_a^{i,N}(a,b] + \bar{\xi}_a^{i,N}(a,b] + \sum_{j=1}^K \left(\bar{\gamma}_b^{ji,N}(a,b] - \bar{\gamma}_a^{ji,N}(a,b]\right).$$

Partition $(\tau, \tau+\delta]$ into $M \in \mathbb{N}$ subintervals $I_m = (t_{m-1}, t_m]$, with $\delta_M = M^{-1}\delta$ and $t_m = \tau + m\delta_M$, $m = 1, ..., M$, and bound the increment of $\bar{\rho}^{i,N}$ by

$$\bar{\rho}^{i,N}(\tau+\delta) - \bar{\rho}^{i,N}(\tau) = \sum_{m=1}^M \left(\bar{\rho}^{i,N}(t_m) - \bar{\rho}^{i,N}(t_{m-1})\right) \leq C_{N,M}^i + D_{N,M}^i + G_{N,M}^i,$$

where

$$\begin{aligned}
C_{N,M}^i &= \sum_{m=1}^M \bar{\xi}_{t_{m-1}}^{i,N}(I_m), \\
D_{N,M}^i &= \sum_{m=1}^M \left(\bar{\alpha}_{t_m}^{i,N}(I_m) - \bar{\alpha}_{t_{m-1}}^{i,N}(I_m)\right), \\
G_{N,M}^i &= \sum_{m=1}^M \sum_{j=1}^K \left(\bar{\gamma}_{t_m}^{ji,N}(I_m) - \bar{\gamma}_{t_{m-1}}^{ji,N}(I_m)\right).
\end{aligned}$$

We first fix M and let $N \to \infty$, and then let $M \to \infty$ to obtain $\mathbb{1}_{E_1^i}\left(C_{N,M}^i + D_{N,M}^i + G_{N,M}^i\right) \to 0$ a.s., which implies $\mathbb{1}_{E_1^i}\left(\rho^i(\tau+\delta) - \rho^i(\tau)\right) = 0$ a.s. This completes the proof by concluding $\mathbb{P}\left(E_1^i \cap E_3^{ic}\right) = 0$ for all i.

By assumption, $D_{N,M}$ converges as $N \to \infty$ to

$$D_M = \sum_{m=1}^{M} \int_{t_{m-1}}^{t_m} a_s^i \left(I_m\right) ds \leq (T+\delta) \sup_{s\in[0,T]} a_s^i \left[s, s+\delta_M\right],$$

and, as $M \to \infty$, $\sup_{s\in[0,T]} a_s^i \left[s, s+\delta_M\right] \to 0$ as assumed in Assumption 1.

Next, note that $C_{N,M}^i \leq M \max_{s\in[\tau,\tau+\delta]} \bar{\xi}_s^{i,N} \left(\tau, \tau+\delta\right]$, and recall that $\bar{\xi}^{i,N} \to \xi^i \in \mathbb{C}_{\mathcal{M}_\sim}$ a.s. Hence

$$\sup_{s\in[0,T]} \sup_x \left|\bar{\xi}_s^{i,N}\left[0,x\right] - \xi_s^i\left[0,x\right]\right| \to 0 \text{ a.s.}$$

This, together with $\mathbb{1}_{E_1^i} \xi_\tau \left[\tau, \tau+\delta\right] = 0$ and $\xi_\tau \left[0,\tau\right) = 0$, implies $\mathbb{1}_{E_1^i} \bar{\xi}_\tau^{i,N} \left[\tau, \tau+\delta\right] \to 0$ a.s. By virtue of the shift property of the Skorokhod mapping, $\xi_{\tau+\cdot}^i \left[0, \tau+\delta\right] = \Gamma_1^{(1)}\left(\psi^{i,\tau,\delta}\right)$, where:

$$\begin{aligned}
\psi^{i,\tau,\delta}\left(t\right) =& \xi_\tau^i \left[0,\tau+\delta\right] + \alpha_{\tau+t}^i \left[0,\tau+\delta\right] - \alpha_\tau^i \left[0,\tau+\delta\right] \\
&+ \sum_{j=1}^{K} P_{ji} \left(\beta_{\tau+t}^{js} \left[0,\tau+\delta-\varepsilon\right] - \beta_\tau^{js} \left[0,\tau+\delta-\varepsilon\right]\right) \\
&- \beta_{\tau+t}^{ir} \left[0,\tau+\delta\right] - \beta_\tau^{ir} \left[0,\tau+\delta\right] - \mu^i\left(\tau+t\right) + \mu^i\left(\tau\right).
\end{aligned}$$

Notice that if δ is smaller than ε then the sum over j is just 0, as in (52) . Moreover,

$$\alpha_{\tau+t}^i \left[0,\tau+\delta\right] - \alpha_\tau^i \left[0,\tau+\delta\right] - \mu^i\left(\tau+t\right) + \mu^i\left(\tau\right) = \int_\tau^{\tau+t} a_s^i \left[0,\tau+\delta\right] ds - \int_\tau^{\tau+t} m^i\left(s\right) ds$$

is non-increasing for $t \in [0,\delta_0]$ by (59). Therefore, $\xi_t^{i,N}\left[0,\tau+\delta\right] = 0$ for all $t \in [\tau,\tau+\delta]$, resulting with $\mathbb{1}_{E_1^i} C_{N,M} \to 0$ a.s.

As for $G_{N,M}^i$, we start with the bound

$$G_{N,M}^i \leq \sum_{m=1}^{M}\sum_{j=1}^{K} \left(\bar{\gamma}_{t_m}^{j,N}\left(I_m\right) - \bar{\gamma}_{t_{m-1}}^{j,N}\left(I_m\right)\right) \leq \sum_{m=1}^{M}\sum_{j=1}^{K} \left(\bar{\beta}_{t_m}^{js,N}\left(I_m\right) - \bar{\beta}_{t_{m-1}}^{js,N}\left(I_m\right)\right) + \frac{MK}{N}.$$

To bound it further, we prove the following. Fix $t_2 > t_1$, then for any interval $A = [t_3, t_4) \subset [0,T]$ such that $t_2 \geq t_4$:

$$\sum_{i=1}^{K} \left(\bar{\beta}_{t_2}^{is,N}\left(A\right) - \bar{\beta}_{t_1}^{is,N}\left(A\right)\right) \leq \sum_{n=0}^{\lfloor t_4/\epsilon\rfloor} \left(\sum_{i=1}^{K} \left(\bar{\alpha}_{t_2}^{i,N}\left(A-n\epsilon\right) - \bar{\alpha}_{t_1}^{i,N}\left(A-n\epsilon\right)\right) + \sum_{i=1}^{K} \bar{\xi}_{t_1}^{i,N}\left(A-n\epsilon\right) + \frac{K}{N}\right)$$

This will be shown by induction over $t_4 < n\epsilon$. From (39) and the monotonicity of $t \mapsto \bar{\beta}_t^{ir,N}\left(A\right)$, and $\bar{\xi}_{t_2}^{i,N}\left[0,t_4\right) = 0$:

$$\bar{\beta}_{t_2}^{is,N}\left(A\right) - \bar{\beta}_{t_1}^{is,N}\left(A\right) \leq \bar{\alpha}_{t_2}^{i,N}\left(A\right) - \bar{\alpha}_{t_1}^{i,N}\left(A\right) + \sum_{j=1}^{K} \left(\bar{\gamma}_{t_2}^{ji,N}\left(A\right) - \bar{\gamma}_{t_1}^{ji,N}\left(A\right)\right) + \bar{\xi}_{t_1}^{i,N}\left(A\right),$$

and by summing over all servers

$$\sum_{i=1}^{K}\left(\bar{\beta}_{t_2}^{is,N}(A)-\bar{\beta}_{t_1}^{is,N}(A)\right)\leq\sum_{i=1}^{K}\left(\bar{\alpha}_{t_2}^{i,N}(A)-\bar{\alpha}_{t_1}^{i,N}(A)+\bar{\xi}_{t_1}^{i,N}(A)\right)$$
$$+\sum_{j=1}^{K}\left(\bar{\beta}_{t_2}^{js,N}(A-\epsilon)-\bar{\beta}_{t_1}^{js,N}(A-\epsilon)\right)+\frac{K}{N}.$$

If $t_4<\epsilon$ then

$$\bar{\beta}_{t_2}^{is,N}(A)-\bar{\beta}_{t_1}^{is,N}(A)\leq\bar{\alpha}_{t_2}^{i,N}(A)-\bar{\alpha}_{t_1}^{i,N}(A)+\bar{\xi}_{t_1}^{i,N}(A),$$

and the statement is true by summing over all servers. Assume that the statement is true for $t_4\leq n\epsilon$. If we take some $t_4\leq(n+1)\epsilon$ and use our induction assumption we get:

$$\begin{aligned}\sum_{i=1}^{K}\left(\bar{\beta}_{t_2}^{is,N}(A)-\bar{\beta}_{t_1}^{is,N}(A)\right)&\leq\sum_{i=1}^{K}\left(\bar{\alpha}_{t_2}^{i,N}(A)-\bar{\alpha}_{t_1}^{i,N}(A)+\bar{\xi}_{t_1}^{i,N}(A)\right)\\&+\sum_{n=1}^{\lfloor t_4/\epsilon\rfloor}\sum_{i=1}^{K}\left(\bar{\alpha}_{t_2}^{i,N}(A-n\epsilon)-\bar{\alpha}_{t_1}^{i,N}(A-n\epsilon)+\bar{\xi}_{t_1}^{i,N}(A-n\epsilon)+\frac{1}{N}\right)+\frac{K}{N}\\&=\sum_{n=0}^{\lfloor t_4/\epsilon\rfloor}\sum_{i=1}^{K}\left(\bar{\alpha}_{t_2}^{i,N}(A-n\epsilon)-\bar{\alpha}_{t_1}^{i,N}(A-n\epsilon)+\bar{\xi}_{t_1}^{i,N}(A-n\epsilon)+\frac{1}{N}\right)\end{aligned}$$

We need this result for $t_1=t_3=t_{m-1}$ and $t_2=t_4=t_m$; if M is big enough then $\bar{\xi}_{t_{m-1}}^{i,N}(A-n\epsilon)=0$ for $n\geq1$, and then:

$$\sum_{i=1}^{K}\left(\bar{\beta}_{t_2}^{is,N}(A)-\bar{\beta}_{t_1}^{is,N}(A)\right)\leq\sum_{n=0}^{\lfloor t_4/\epsilon\rfloor}\sum_{i=1}^{K}\left(\bar{\alpha}_{t_2}^{i,N}(A-n\epsilon)-\bar{\alpha}_{t_1}^{i,N}(A-n\epsilon)+\frac{1}{N}\right)+\sum_{i=1}^{K}\bar{\xi}_{t_1}^{i,N}(A).$$

It follows that

$$G_{N,M}^{i}\leq\sum_{m=1}^{M}\frac{KT}{\epsilon}\max_{i,n}\left(\bar{\alpha}_{t_m}^{i,N}(I_m-n\epsilon)-\bar{\alpha}_{t_{m-1}}^{i,N}(I_m-n\epsilon)\right)+K\sum_{m=1}^{M}\bar{\xi}_{t_{m-1}}^{i,N}(I_m)+\left(\frac{T}{\epsilon}+1\right)\frac{MK}{N}.$$

The terms on the RHS vanish as $C_{N,M}^{i}$ and $D_{N,M}^{i}$ above. □

Acknowledgement. This research was supported in part by the ISF (grants 1184/16 and 1035/20).

References

[1] C. M. Aras, J. F. Kurose, D. S. Reeves, and H. Schulzrinne. Real-time communication in packet-switched networks. *Proceedings of the IEEE*, 82(1):122–139, Jan 1994.

[2] R. Atar, A. Biswas, and H. Kaspi. Fluid limits of G/G/1+G queues under the nonpreemptive earliest-deadline-first discipline. *Mathematics of Operations Research*, 40(3):683–702, 2015.

[3] R. Atar, A. Biswas, and H. Kaspi. Law of large numbers for the many-server earliest-deadline-first queue. *Stochastic Processes and their Applications*, 128(7):2270–2296, 2018.

[4] R. Atar, A. Biswas, H. Kaspi, and K. Ramanan. A Skorokhod map on measure-valued paths with applications to priority queues. *The Annals of Applied Probability*, 28(1):418–481, 02 2018.

[5] R. Atar and P. Dupuis. Large deviations and queueing networks: methods for rate function identification. *Stochastic processes and their applications*, 84(2):255–296, 1999.

[6] R. Atar, H. Kaspi, and N. Shimkin. Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research*, 39(3):672–696, 2014.

[7] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9. x+277 pp. A Wiley-Interscience Publication.

[8] M. Bramson. Stability of earliest-due-date, first-served queueing networks. *Queueing Systems*, 39(1):79–102, Sep 2001.

[9] G. C. Buttazzo. *Hard real-time computing systems: predictable scheduling algorithms and applications*, volume 24. Springer Science & Business Media, 2011.

[10] J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.

[11] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks*. Springer-Verlag New York, 2001.

[12] J. G. Dai. On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*, pages 49–77, 1995.

[13] L. Decreusefond and P. Moyal. Fluid limit of a heavily loaded EDF queue with impatient customers. *Markov Processes and Related Fields*, 14:131–157, 2008.

[14] B. Doytchinov, J. Lehoczky, and S. Shreve. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *The Annals of Applied Probability*, 11(2):332–378, 2001.

[15] H. C. Gromoll, A. L. Puha, and R. J. Williams. The fluid limit of a heavily loaded processor sharing queue. *The Annals of Applied Probability*, 12(3):797–859, 2002.

[16] J. M. Harrison and M. I. Reiman. Reflected Brownian motion on an orthant. *The Annals of Probability*, 9(2):302–308, 1981.

[17] C. Heyde and P. Hall. *Martingale Limit Theory and its Application*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, 1980. ix - xi pp.

[18] X. Hu, S. Barnes, and B. Golden. Applying queueing theory to the study of emergency department operations: A survey and a discussion of comparable simulation studies. *International Transactions in Operational Research*, 25, 03 2017.

[19] L. Kruk. Stability of two families of real-time queueing networks. *Probability and Mathematical Statistics*, 28, 01 2008.

[20] L. Kruk. Invariant states for fluid models of EDF networks: Nonlinear lifting map. *Probability and Mathematical Statistics*, 30(2):289–315, 2010.

[21] L. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Double Skorokhod map and reneging real-time queues. *Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz*, Volume 4:169–193, 2008.

[22] L. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Heavy traffic analysis for EDF queues with reneging. *The Annals of Applied Probability*, 21(2):484–545, 04 2011.

[23] L. Kruk, J. Lehoczky, S. Shreve, and S.-N. Yeung. Earliest-deadline-first service in heavy-traffic acyclic networks. *The Annals of Applied Probability*, 14(3):1306–1352, 08 2004.

[24] J. P. Lehoczky. Real-time queueing theory. In *Proceedings of the 17th IEEE Real-Time Systems Symposium*, RTSS '96, pages 186–, Washington, DC, USA, 1996. IEEE Computer Society. ISBN 0-8186-7689-2.

[25] J. P. Lehoczky. Real-time queueing network theory. In *Proceedings Real-Time Systems Symposium*, pages 58–67, Dec 1997.

[26] P. Moyal. On queues with impatience: stability, and the optimality of earliest deadline first. *Queueing Systems*, 75(2): 211–242, Nov 2013.

[27] S. Panwar, D. Towsley, and J. Wolf. Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. *Journal of the ACM*, 35(4):832–844, 10 1988.

[28] S. Ramasubramanian. A subsidy-surplus model and the Skorokhod problem in an orthant. *Mathematics of Operations Research*, 25(3):509–538, 2000.

[29] M. I. Reiman. Open queueing networks in heavy traffic. *Mathematics of operations research*, 9(3):441–458, 1984.