A Skorokhod Map on Measure-Valued Paths with Applications to Priority Queues

Rami Atar[*] Anup Biswas[†] Haya Kaspi[‡] Kavita Ramanan[§]

Abstract

The Skorokhod map on the half-line has proved to be a useful tool for studying processes with non-negativity constraints. In this work we introduce a measure-valued analog of this map that transforms each element ζ of a certain class of càdlàg paths that take values in the space of signed measures on $[0,\infty)$ to a càdlàg path that takes values in the space of non-negative measures on $[0,\infty)$ in such a way that for each $x>0$, the path $t\mapsto\zeta_t[0,x]$ is transformed via a Skorokhod map on the half-line, and the regulating functions for different $x>0$ are coupled. We establish regularity properties of this map and show that the map provides a convenient tool for studying queueing systems in which tasks are prioritized according to a continuous parameter. Three such well known models are the *earliest-deadline-first*, the *shortest-job-first* and the *shortest-remaining-processing-time* scheduling policies. For these applications, we show how the map provides a unified framework within which to form fluid model equations, prove uniqueness of solutions to these equations and establish convergence of scaled state processes to the fluid model. In particular, for these models, we obtain new convergence results in time-inhomogeneous settings, which appear to fall outside the purview of existing approaches.

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

Keywords: Skorokhod map, measure-valued Skorokhod map, measure-valued processes, fluid models, fluid limits, law of large numbers, priority queueing, Earliest-Deadline-First, Shortest-Remaining-Processing Time, Shortest-Job-First.

Contents

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

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

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

[§]Division of Applied Mathematics, Brown University, Providence, RI 02118, USA

1 Introduction

An established framework in queueing theory is to identify scaling limits of system dynamics, whereby one can describe the qualitative behaviour of processes such as queue length, workload and other performance measures. In this context, the classical Skorokhod map (SM) introduced by Skorokhod [34] and its multi-dimensional analogs, have served as useful tools for establishing limit theorems. The classical Skorokhod map, which in this paper we refer to as the SM on the half-line, acts on real-valued paths to produce a path that is constrained to be non-negative. By representing the queue-length process as the image, under a (possibly multi-dimensional) SM, of a simpler so-called "netput" process, one can often reduce the problem of establishing convergence of the sequence of queue-length processes to the simpler problem of establishing convergence of the corresponding sequence of netput processes. In recent years, the study of more complex networks has led to the use of measure-valued processes, which have proved powerful for analyzing both single-server and many-server systems, and specifically, establishing Law of Large Numbers (LLN) and Central Limit Theorem (CLT) results [3, 4, 9, 12, 13, 15, 19, 20, 21, 26, 30, 36]. In the context of single-server networks, measure-valued processes have been particularly useful for studying scaling limits of models in which jobs are prioritized according to a continuous parameter, such as the deadline of the job or the job size. In this case, the measures for which the dynamics are specified correspond to the (suitably normalized) counting measure that keeps track of the number of jobs with deadlines and sizes, respectively, in any given interval. In this work, we introduce a map acting on a subset of paths in the space of signed measures that can be viewed as a measure-valued analog of the classical SM, and which we refer to as the measure-valued Skorokhod map (MVSM). We show that the MVSM provides a unified framework for the study of the dynamics of queueing systems with continuous parameter priority scheduling policies. Specifically, we use the map to formulate fluid models of several such queueing systems, as well as to prove LLN results for these systems, both in time-homogeneous and time-varying settings. The map and its regularity properties may be of independent interest and could potentially also have applications in other fields.

To describe the MVSM, let $\mathcal{M}$ and $\mathcal{M}'$ denote the spaces of finite non-negative measures and signed measures, respectively, on the non-negative real line. Given (α, μ), where α is an $\mathcal{M}$-valued càdlàg path and μ is a non-negative non-decreasing càdlàg real-valued function on $[0, \infty)$, the MVSM maps the $\mathcal{M}'$-valued path $t \mapsto \alpha_t - \mu(t)$ to an $\mathcal{M}$-valued càdlàg path in such a way that for each $x \geq 0$, the real-valued path $t \mapsto \alpha_t[0, x] - \mu(t)$ is transformed under the classical SM on the half-line and the constraining terms for different x are coupled in a specific fashion (see Definition 2.5 for a precise description). Our key observation is that the MVSM serves as a generic model for priority. We demonstrate this point by applying the MVSM to study several queueing models employing a continuous parameter priority that have been previously treated by distinct tools, and to obtain new results for models that seem to fall outside the purview of existing methods.

Among the several scheduling policies for which we argue that the MVSM is applicable, we treat three in detail: Earliest-Deadline-First (EDF), Shortest-Remaining-Processing-Time (SRPT) and Shortest-Job-First (SJF). In EDF, jobs are prioritized according to their deadlines, which are declared upon arrival. We consider two versions of the policy, depending on whether the jobs are subject to "soft" or "hard" deadlines. If jobs continue to be accepted into service even after their deadlines have elapsed, then we refer to this as the "soft EDF" policy, whereas with the "hard EDF" policy, jobs that miss their deadlines renege (are ejected from the system). The soft or hard EDF policy is said to be preemptive if an arriving job with a more urgent deadline is allowed to interrupt a job in service, and non-preemptive otherwise. We have chosen to treat here only the non-preemptive policy, for both the soft and hard versions of EDF. In the SRPT and SJF policies, scheduling is prioritized according to the size of a job (for a survey and motivation regarding these policies we refer to the introduction in [8]). Under SRPT, the arrival of a job whose size is smaller than the remaining service time of the one being currently processed will interrupt the service, whereas service is non-interruptible under SJF. In other words, SRPT and SJF are, respectively, preemptive and non-preemptive versions of a common priority policy.

To set our results in context, we first discuss prior work on the EDF, SRPT and SJF models. The EDF model was first considered in [9] as far as scaling limits are concerned, and further results appeared in [26]. In both papers diffusion approximations in heavy traffic were established; [9] treats the preemptive soft EDF model with general renewal arrivals and independent and identically distributed (i.i.d.) service times (the so-called $GI/GI/1$ setting), whereas [26] analyzes the preemptive hard EDF (or $GI/GI/1+GI$) version of the model, with both works considering jobs that have i.i.d. deadlines drawn from a general distribution. The analysis in [26] is carried out by introducing a map (see Section 4.1.1 therein) that transforms the space of càdlàg $\mathcal{M}$-valued paths to itself in such a way that it acts on the measure-valued state process of the preemptive soft EDF model to obtain an approximation of the corresponding state process in the preemptive hard EDF model, which becomes exact in the heavy-traffic limit. As elaborated in [26, 25], this map can be viewed as a measure-valued generalization of the map on real-valued paths that takes the image of the SM on the half-line to the image of the so-called double-barrier SM on a bounded interval $[0, a]$ [24], and thus, is completely different from our MVSM. In terms of LLN limits, the non-preemptive hard EDF model was studied in [7] and [2]. The former considered general deadline distributions but Poisson arrivals and exponential service times (the $M/M/1 + GI$ setting) by analyzing the Markov evolution of a measure-valued state process, whereas the latter considered the case of general arrivals and service times with general deadline distributions (the $G/GI/1+GI$ setting) that satisfy a certain monotonicity condition, and made key use of a certain Skorokhod problem with a time-varying barrier. All the existing LLN and CLT results for the (soft or hard) EDF policy mentioned above [2, 7, 9, 26] heavily rely on the assumption that the arrival and service rates are constant, and more specifically, crucially use the so-called *frontier* process, a concept that was introduced in [9]. The frontier at time t is defined to be the maximum of the *lead* *times* of jobs present in the system at time t that have ever been in service (here, the lead time of a job is defined to be its deadline minus the current time). The results crucially rely on the fact that under suitable conditions, the asymptotic behavior is such that the frontier process separates the population of jobs into those that have been sent to service and those that have not. However, such a frontier process may not exist in general. In particular, as illustrated in Figure 1.1 and supported by computer simulations, there is typically no such separation of populations when the arrival or service rate is time varying.

As for results on scaling limits of the SRPT and SJF policies, CLTs for queues in heavy traffic working under the SRPT policy have been established in [13, 30], while LLN results for the SRPT and SJF policies have been established in [8] and [14]. As shown in [14], the limits under both policies agree. As in the prior work on EDF, the works [14] and [8] also make use of an analogously defined frontier process, and assume constant arrival and service rates.

In this paper we apply the common framework of the MVSM to establish LLN results for the EDF, SRPT and SJF policies, in particular allowing for time-inhomogeneous arrival and service rates. For the EDF policy, we use the term *patience* to denote the time that an arriving job is willing to wait in queue before it reneges. Thus the deadline of any given job is the sum of its arrival time and its patience (as elaborated in Section 4, we use the notion of *absolute* deadlines, as opposed to *relative* deadlines, defined relative to the current time; note that another term for the latter is the lead time, already mentioned above). We establish the LLN limit of a queue operating under the non-preemptive hard EDF policy, in which jobs with patience that follows a general, possibly time-inhomogeneous distribution, arrive to a single-server queue and the cumulative arrival and service processes are modelled by general, possibly time-inhomogeneous non-decreasing stochastic processes. The result we obtain is far more general than [2] and [7] as it allows variable arrival and service rates and also relaxes the assumption made in [2] regarding strict monotonicity of the deadline distribution function. Moreover, the treatment of the fluid model equations establishes a result that may be of independent interest, which shows that EDF scheduling is optimal in the LLN limit in terms of the reneging count. Earlier results on this aspect include [28, 29], where the optimality of EDF, in terms of minimizing the total number of reneged jobs, is shown for the $G/M/1+GI$ queue. In [26] it is shown that the total amount of *reneged work* is optimized in a $G/G/1+G$ queue when the EDF scheduling policy is applied. Optimality properties of EDF are also studied in [27]. For the SJF and SRPT policies, we generalize the results in [8, 14] to allow time-varying arrival and service rates, where again, the notion of a frontier becomes ineffective. Also, our proof technique, which involves the application of the MVSM in conjunction with the continuous mapping theorem, substantially simplifies the analysis.

Although we consider the performance of priority policies at a single queue in this paper, we believe that a suitable extension of the MVSM approach could also be useful for the study of networks. Past results regarding the soft EDF policy in a network context are as follows. Queueing networks with random routing under the soft EDF policy without preemption were studied in [5] (referred to there as earliest-due-date-first-served), where it was shown that subcritical networks are stable by analyzing the associated fluid model. This result was extended in [22] to the case of preemptive subcritical EDF networks when job routes are fixed by studying the fluid model and showing that it satisfies the FIFO (first-in-first-out) fluid model equations. This work also established a stability theorem for a broader class of (not necessary subcritical) networks with reneging, but without recourse to fluid model equations. The main idea in [22] is to show that the initial lead time distribution vanishes in the limit, and thus EDF reduces to FIFO. With a view to extending the general theory of heavy-traffic limits for multiclass queueing networks to a class of non-head-of-the-line scheduling policies, the paper [23] also studies fluid limits of EDF networks and characterizes its invariant states. The MVSM approach could potentially be useful for obtaining results for SRPT and SJF networks, where there are not many existing results.

It is worth pointing out that a completely different extension of the classical SM that acts on measure-valued paths (or more general real-valued functions defined on a poset) was considered in [1]. However, while interesting on its own, when applied to our setting this extension provides a decomposition that is not useful for the applications considered here (see Remark 2.11).

To summarize our main contributions in this paper, we have

- Introduced and established regularity properties of a Skorokhod-type map, the MVSM, that acts on a space of measure-valued paths;
- Shown that this map serves as a natural tool for analyzing priority queueing models with continuous parameter, and used the map to formulate fluid models for (both hard and soft) EDF, SJF and SRPT;
- Developed a unified method for establishing LLN limits for the aforementioned policies, including in time-inhomogeneous situations in which the notion of a frontier, which was used in previous analyses, is ineffective.

In addition to the time-inhomogeneous case being of intrinsic interest since it is often the generic situation in applications, another motivation for our analysis is that the MVSM is likely to also be pertinent for the study of (even time-homogeneous) many-server systems with general service and deadline distributions operating under the EDF, SRPT or SJF policies. Moreover, we believe that this approach, and in particular the MVSM, will also be useful for the analysis of other queueing models in which there is prioritization with respect to a continuous parameter (such as, e.g., [35, 32]). Furthermore, the MVSM, or its close relative, may potentially also be useful for the study of interacting particle systems arising in other fields. Such applications will be explored in future work.

The organization of the paper is as follows. First, in Section 1.1 we collect some common notation used in the paper. In Section 2 we introduce the measure-valued Skorokhod problem (MVSP), which defines the MVSM, and establish properties of the map. In Section 3 we introduce some illustrative examples that serve to motivate the form of the MVSP. In Section 4 and Section 5 we describe fluid models and establish LLN results, respectively, for the EDF, SJF and SRPT policies: Sections 4.1 and 5.1 are devoted to the EDF model, while Sections 4.2 and 5.2 focus on the SJF and SRPT policies.

1.1 Notation

For $x, y \in \mathbb{R}$, the maximum [minimum] is denoted by $x \vee y$ [resp., $x \wedge y$]. For $A \subset \mathbb{R}_+ := [0, \infty)$, define $A^\varepsilon = \{x \geq 0 : \inf_{a \in A} |x - a| < \varepsilon\}$ and let $\inf A$ (respectively, $\min A$) denote the infimum (respectively, minimum, if it exists) of the set of points in A. Denote by $\mathbb{I}_A$ the indicator function of a set A, which takes the value 1 on the set A, and zero otherwise. For $f : \mathbb{R} \to \mathbb{R}^k$ denote $\|f\|_T = \sup_{t\in[0,T]} \|f(t)\|$, and for $\varepsilon > 0$, we define the oscillation of f as

$$Osc_\varepsilon(f) \doteq \sup\{|f(s) - f(t)| : |s - t| \le \varepsilon, s, t \in \mathbb{R}\}.$$

For a topological space $\mathcal{S}$, denote by $\mathbb{C}_b(\mathcal{S})$ the set of real-valued bounded, continuous maps on $\mathcal{S}$, by $\mathbb{C}_{b,+}(\mathcal{S})$ the collection of members of $\mathbb{C}_b(\mathcal{S})$ that are non-negative, and by $\mathcal{B}(\mathcal{S})$ the Borel σ-field on $\mathcal{S}$. For a Polish space $\mathcal{S}$, denote by $\mathbb{C}_\mathcal{S}$ the space of continuous functions $\mathbb{R}_+ \to \mathcal{S}$ and $\mathbb{D}_\mathcal{S}$, the space of functions $\mathbb{R}_+ \to \mathcal{S}$ that are right continuous at every $t \in [0,\infty)$ and have finite left limits at every $t \in (0,\infty)$. The space $\mathbb{D}_\mathcal{S}$ is endowed with the Skorohod J_1 topology and $\mathbb{C}_\mathcal{S}$ is endowed with the topology of uniform convergence on compact subsets. Also, let $\mathbb{D}^\uparrow_\mathbb{R}$ (respectively, $\mathbb{C}^\uparrow_\mathbb{R}$) denote the subset of functions in $\mathbb{D}_\mathbb{R}$ that are non-negative and non-decreasing.

The space of non-negative finite Borel measures on $\mathbb{R}_+$ is denoted by $\mathcal{M}$, and the subspace of measures in $\mathcal{M}$ that have no atoms are denoted by $\mathcal{M}_0$. Given $\nu \in \mathcal{M}$, we let supp$[\nu]$ denote the support of ν, which is defined to be the closure of the set of points $x \in \mathbb{R}_+$ for which every open neighborhood N_x of x has positive measure, that is, $\nu(N_x) > 0$. Given two measures $\nu, \nu' \in \mathcal{M}$, we will write $\nu \ll \nu'$ to denote that ν is absolutely continuous with respect to ν'. The symbol δ_x denotes the point mass at $x \in \mathbb{R}_+$. For $\nu \in \mathcal{M}$ and a Borel measurable function g on $\mathbb{R}_+$, we use the notation $\langle g, \nu\rangle = \int g d\nu$. Endow $\mathcal{M}$ with the Levy metric given by

$$d_\mathcal{L}(\nu_1, \nu_2) = \inf\{\varepsilon > 0 : \nu_1[0,(x-\varepsilon)^+] - \varepsilon \le \nu_2[0,x] \le \nu_1[0,x+\varepsilon] + \varepsilon, \text{for all } x \in \mathbb{R}_+\}. \quad (1.1)$$

It is well known that $(\mathcal{M}, d_\mathcal{L})$ is a Polish space [16, Chapter 2]. Also, the topology induced by $d_\mathcal{L}$ is equivalent to the weak topology on $\mathcal{M}$, characterized by $\nu_n \to \nu$ in $\mathcal{M}$ if and only if

$$\langle f, \nu_n\rangle \to \langle f, \nu\rangle \text{ for all } f \in \mathbb{C}_b(\mathbb{R}_+).$$

For $\nu \in \mathcal{M}$, we write $\nu[a,b]$ for $\nu([a,b])$, and similarly $\nu[a,b)$ for $\nu([a,b))$, etc. It is well known that

$$d_\mathcal{L}(\nu_1,\nu_2) \le \sup_{x\in[0,\infty)} |\nu_1[0,x] - \nu_2[0,x]| \le d_\mathcal{L}(\nu_1,\nu_2) + Osc_{2d_\mathcal{L}(\nu_1,\nu_2)}(\nu_2[0,\cdot]) \quad (1.2)$$

(for the first inequality see [16, Eq. (2.25)], the second follows by definition). On the other hand, given $\xi \in \mathbb{D}_\mathcal{M}$ and $0 \le a \le b$, we use $\xi.[a,b]$ to denote the function $t \mapsto \xi_t[a,b]$. Also, given $t \ge 0$, if $\zeta \in \mathbb{D}_\mathcal{M}$ we will use ζ_t to denote the evaluation of the path ζ at time t, whereas if $f \in \mathbb{D}_\mathbb{R}$, then we will use $f(t)$ to denote the value of f at time t.

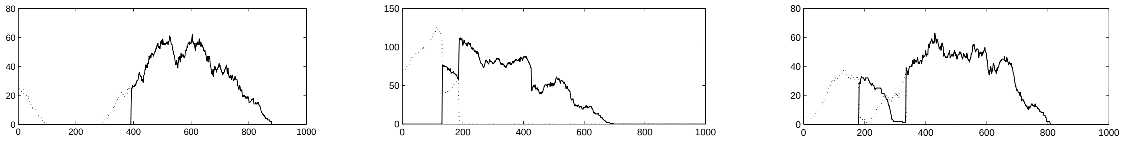
For $\zeta \in \mathbb{D}^\uparrow_\mathbb{R}$ we denote by γ^ζ the Lebesgue-Stieltjes measure that ζ induces on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, namely,

$$\gamma^\zeta(B) = \zeta(0)\delta_0(B) + \int_{(0,\infty)} \mathbb{I}_B(t) d\zeta_t, \quad B \in \mathcal{B}(\mathbb{R}_+). \quad (1.3)$$

Throughout, we write "$d\zeta$-a.e." to mean "$d\gamma^\zeta$-a.e."

2 A Skorokhod problem on the space of measure-valued paths

In Section 2.1 we recall the definition of the Skorokhod problem (SP) on the half-line, and list some properties that will be useful in our analysis. In Section 2.2 we introduce the MVSM.



(a) (b) (c)

Figure 1: *Simulation results for the hard EDF model in which the arrival stream is stochastic and highly inhomogeneous. The graphs depict histograms of number of jobs as a function of the lead time (i.e., the time until a job's deadline), at three different epochs. Jobs that have arrived into the system and have [have not] been sent to service are shown in dotted [resp., solid] line. At the epoch captured in graph (a), these two populations of jobs are separated by a certain priority level, often referred to as the frontier. Graphs (b) and (c) correspond to a generic situation, in which the notion of a frontier is no longer effective. The service rate is fixed at 60 jobs per unit time, while the arrival pattern changes periodically (with 400 time units period) between a uniform distribution over* $[50, 299]$ *at rate 100 and* $[600, 849]$ *at rate 50 jobs per unit time.*

2.1 The Skorokhod Map on the Half-Line

The SP on the half-line was first introduced by Skorokhod in [34]. Roughly speaking, it seeks to transform a real-valued function to one that is minimally constrained to be non-negative.

Definition 2.1 (Skorokhod problem (SP) on the half-line) *Given data* $\psi \in \mathbb{D}_{\mathbb{R}}$*, find a pair* $(\varphi, \eta) \in \mathbb{D}_{\mathbb{R}} \times \mathbb{D}_{\mathbb{R}}^{\uparrow}$ *such that* $\varphi = \psi + \eta$*,* $\varphi(t) \geq 0$ *for all* $t \geq 0$*, and* $\varphi(t) = 0$ *for* $d\eta$*-a.e.* $t \in [0, \infty)$*.*

It is well known that for every $\psi \in \mathbb{D}_{\mathbb{R}}$, there is a unique solution $(\varphi, \eta) = \Gamma[\psi]$ that solves the SP on the half-line, and we refer to Γ as the *Skorokhod map (SM) on the half-line*. Specifically, if we denote the two component maps of Γ by Γ_1 and Γ_2, then for $\psi \in \mathbb{D}_{\mathbb{R}}$,

$$\varphi(t) =: \Gamma_1[\psi](t) = \psi(t) - \inf_{s \in [0,t]} (\psi(s) \wedge 0), \qquad \eta(t) =: \Gamma_2[\psi](t) = \varphi(t) - \psi(t), \qquad t \geq 0. \quad (2.1)$$

We now state two elementary properties of the SM Γ.

Lemma 2.2 *For* $i = 1, 2$*, let* $\psi_i \in \mathbb{D}_{\mathbb{R}}$ *and* $(\varphi_i, \eta_i) = \Gamma[\psi_i]$*. Then the following properties hold.*

1. *(Monotonicity) If* $\psi_2 - \psi_1 \in \mathbb{D}_{\mathbb{R}}^{\uparrow}$ *then* $\eta_1 - \eta_2 \in \mathbb{D}_{\mathbb{R}}^{\uparrow}$ *and* $\varphi_2 - \varphi_1 \geq 0$*.*
2. *(Lipschitz continuity)* $\|\varphi_2 - \varphi_1\|_T \leq 2\|\psi_2 - \psi_1\|_T$ *for any* $T \in (0, \infty)$*.*

Proof: The statements follow immediately from the explicit formula for (φ, η) in (2.1). □

We close this section by stating two more basic properties of Γ_1 that will be used frequently in the sequel. In what follows, given a real-valued function f on $[0, \infty)$ and $T > 0$, we define the shifted version f^T as follows:

$$f^T(t) = f(T + t) - f(T), t \geq 0. \quad (2.2)$$

Lemma 2.3 *Given $\psi \in \mathbb{D}_{\mathbb{R}}$, let $\varphi = \Gamma_1(\psi)$. Then the following two properties hold:*

1. *Given $T \in [0, \infty)$, $\varphi(T + t) = \Gamma_1(\varphi(T) + \psi^T)(t)$, $t \geq 0$.*

2. *For any $0 \leq S \leq T < \infty$, $\varphi(T) = 0$ if and only if $\psi^S(T - S) = \inf_{s\in[0,T-S]} \psi^S(s) \leq -\varphi(S)$. Moreover, if $\varphi(S) = 0$, then $\varphi(T) = 0$ for $T \in [S, S + \delta]$ if and only if ψ is non-increasing on $[S, S + \delta]$.*

Proof: The first property is easy to verify directly from the properties of the SP (see also Lemma 2.3 of [31]). Moreover, for $0 \leq S \leq T < \infty$ property 1. and (2.1) imply that

$$\varphi(T) = \varphi(S) + \psi^S(T - S) - \inf_{s\in[0,T-S]} \left((\varphi(S) + \psi^S(s)) \wedge 0\right).$$

Property 2 is a simple consequence of this relation. □

2.2 The MVSP: definition and properties

In this section, we define a measure-valued Skorohod problem (MVSP), show that it possesses a unique solution, and refer to the solution map as the measure-valued Skorokhod map (MVSM). We then establish certain regularity properties of this map. To this end, let

$$\mathbb{D}^{\uparrow}_{\mathcal{M}} := \{\zeta \in \mathbb{D}_{\mathcal{M}} : t \mapsto \langle f, \zeta_t\rangle \text{ is non-decreasing } \forall f \in \mathbb{C}_{b,+}(\mathbb{R}_+)\}, \tag{2.3}$$

where recall that $\mathbb{C}_{b,+}(\mathbb{R}_+)$ is the space of non-negative bounded continuous maps on $\mathbb{R}_+$. The following lemma gathers some elementary properties of the space $\mathbb{D}^{\uparrow}_{\mathcal{M}}$. Its proof is relegated to Appendix A.

Lemma 2.4 *The following properties hold.*

1. *$\mathbb{D}^{\uparrow}_{\mathcal{M}}$ is a closed subset of $\mathbb{D}_{\mathcal{M}}$.*

2. *If $\zeta \in \mathbb{D}_{\mathcal{M}}$, then $\zeta \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$ if and only if for every $0 \leq x < y$, $\zeta[0, x] \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ and $\zeta(x, y] \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$.*

3. *If $t \mapsto \zeta_t$ is in $\mathbb{D}^{\uparrow}_{\mathcal{M}}$, then for every $t, x \in \mathbb{R}_+$ and sequences $\{x_n\}, \{y_n\} \subset \mathbb{R}_+$ such that $x_n \downarrow x$, and $y_n \uparrow x$,*

$$\lim_{n\to\infty} \sup_{s\in[0,t]} \zeta_s(x, x_n] = 0 \quad \textit{and} \quad \lim_{n\to\infty} \sup_{s\in[0,t]} \zeta_s(y_n, x) = 0. \tag{2.4}$$

4. *Given any measurable space $(S, \mathcal{S})$, a map $\mathcal{T}$ from $(S, \mathcal{S})$ to $\mathbb{D}_{\mathcal{M}}$, equipped with the Borel σ-algebra, is measurable if and only if for every $t, x \geq 0$, the map $\mathcal{T}_{t,x} : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{T}_{t,x}(s) = (\mathcal{T}(s))_t[0, x]$, is measurable.*

We now define a solution (ξ, β, ι) to the MVSP with data (α, μ). As shown in Sections 3.1 and 3.2, the definition of the MVSP given below can be seen as a natural generalization of the equations (3.5) that describe a K-class priority queue with α modeling the arrivals and μ the service rate to the case when there is a continuum of priority classes. The reader interested primarily in the queueing application may find it useful to read Section 3.1 before looking at Definition 2.5.

Definition 2.5 (MVSP) *Let* $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$*. Then* $(\xi, \beta, \iota) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ *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. $\xi[0, x] = 0$ $d\beta(x, \infty)$*-a.e.,*
3. $\xi[0, x] = 0$ $d\iota$*-a.e.,*
4. $\beta[0, \infty) + \iota = \mu$*.*

Remark 2.6 If $(\xi, \beta, \iota) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ solves the MVSP for the data $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$, then for each fixed t, sending $x \to \infty$ in property 1, we see that

$$\xi_t[0, \infty) = \alpha_t[0, \infty) - \mu + \iota, \qquad t \geq 0. \tag{2.5}$$

properties 1 and 4 of Definition 2.5 imply that for $t \geq 0$, we have the simple balance relation $\xi_t[0, x] = \alpha_t[0, x] - \beta_t[0, x]$ for $x \in \mathbb{R}_+$, and therefore that

$$\xi_t(A) = \alpha_t(A) - \beta_t(A), \quad A \in \mathcal{B}(\mathbb{R}_+). \tag{2.6}$$

In turn, note that (2.6) implies that for every $t \geq 0$,

$$\xi_t \ll \alpha_t \qquad \text{and} \qquad \beta_t \ll \alpha_t, \tag{2.7}$$

where recall that $\nu \ll \nu'$ denotes that ν is absolutely continuous with respect to ν'.

We now establish an alternative characterization of the MVSP in terms of the SP on the half-line, which is useful for establishing uniqueness of a solution to the MVSP .

Lemma 2.7 *Let* $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ *and let* Γ *be the SM on the half-line (see Definition 2.1). Then* $(\xi, \beta, \iota) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ *satisfy properties 1–4 of Definition 2.5 if and only if*

$$\big(\xi[0, x], \beta(x, \infty) + \iota\big) = \Gamma\big[\alpha[0, x] - \mu\big], \qquad x \in \mathbb{R}_+, \tag{2.8}$$

and

$$\beta(\{0\}) = \alpha(\{0\}) - \xi(\{0\}). \tag{2.9}$$

Moreover, if (ξ, β, ι) *satisfy properties 1–4 of Definition 2.5 then*

$$\big(\xi[0, \infty), \iota\big) = \Gamma\big[\alpha[0, \infty) - \mu\big], \tag{2.10}$$

and for every $x \in \mathbb{R}_+$*,*

$$\big(\xi[0, x], \beta(x, \infty)\big) = \Gamma\big[\alpha[0, x] - \mu + \iota\big], \qquad x \in \mathbb{R}_+. \tag{2.11}$$

Proof: First, suppose properties 1–4 of Definition 2.5 are satisfied. Then, properties 2 and 3 of Definition 2.5 are equivalent to the conditions $\xi[0,x] = 0$ $d(\beta(x,\infty)+\iota)$-a.e. for every $x \in \mathbb{R}_+$. By Definition 2.1 of the SM Γ the latter two relations in conjunction with property 1 of Definition 2.5 and (2.5) imply (2.8). Now, property 1 of the MVSP implies that

$$\xi(\{0\}) = \alpha(\{0\}) - \mu + \beta(0,\infty) + \iota. \tag{2.12}$$

Together with property 4 of the MVSP , which can be rewritten as $\beta(0,\infty) = \mu - \iota - \beta(\{0\})$, this implies (2.9). Also, note that (2.8) implies that for every $t, x \in \mathbb{R}_+$, $\xi_t[0,x] = \alpha_t[0,x] - \mu(t) + \beta_t(x,\infty) + \iota(t)$ and that $\xi_t[0,x] = 0$ for $d\iota$ a.e. t. Sending $x \to \infty$ we see that $\xi_t[0,\infty) = \alpha_t[0,\infty) - \mu(t) + \iota(t)$ and $\xi_t[0,\infty) = 0$ for $d\iota$ a.e. t, and therefore, (2.10) follows. On the other hand, (2.11) follows from property 1 of Definition 2.5, the fact that $\xi[0,x] \geq 0$, $\beta(x,\infty) \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ and property 2 of Definition 2.5.

Now, for the converse, suppose (2.8)-(2.9) holds. Then the definition of Γ shows that properties 1 and 3 of Definition 2.9 hold, and also that $\xi_t[0,x] = 0$ $d(\beta(x,\infty)+\iota)$-a.e., which implies properties 2 and 3 since $\beta(x,\infty)$ and ι are both non-decreasing. Now, (2.8) with $x = 0$ implies (2.12), which when combined with (2.9) implies property 4 of the MVSP . This completes the proof of the first assertion of the lemma. □

We now show that the MVSP has a unique solution and preserves certain continuity properties. Analogous to (2.3), we let $\mathbb{C}^{\uparrow}_{\mathbb{R}}$ denote the subset of functions in $\mathbb{C}_{\mathbb{R}}$ that are non-negative and non-decreasing, and define

$$\mathbb{C}^{\uparrow}_{\mathcal{M}} := \{\zeta \in \mathbb{C}_{\mathcal{M}} : \ t \mapsto \langle f, \zeta_t \rangle \text{ is non-decreasing } \forall f \in \mathbb{C}_{b,+}(\mathbb{R}_+)\}, \tag{2.13}$$

and let

$$\mathbb{C}^{\uparrow}_{\mathcal{M}_0} := \{\zeta \in \mathbb{C}^{\uparrow}_{\mathcal{M}} : \text{ for each } t, \ \zeta_t \in \mathcal{M}_0\}, \tag{2.14}$$

where recall that $\mathcal{M}_0 \subset \mathcal{M}$ is the subset of measures in $\mathcal{M}$ that have no atoms.

Proposition 2.8 *For every $(\alpha,\mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ there exists a unique solution $(\xi,\beta,\iota) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ to the MVSP . Moreover, if $\alpha \in \mathbb{D}^{\uparrow}_{\mathcal{M}_0}$ then $(\xi,\beta) \in \mathbb{D}_{\mathcal{M}_0} \times \mathbb{D}^{\uparrow}_{\mathcal{M}_0}$. Further, if $(\alpha,\mu) \in \mathbb{C}^{\uparrow}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$, then the corresponding solution (ξ,β,ι) lies in $\mathbb{C}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$.*

Proof: Fix $(\alpha,\mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$. We first explicitly construct a candidate solution to the MVSP for (α,μ). It follows from (2.10) of Lemma 2.7 that the ι-component of the solution must satisfy

$$\iota := \Gamma_2[\alpha[0,\infty) - \mu]. \tag{2.15}$$

Note that ι thus defined does indeed lie in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$. Next, in view of relation (2.11) of Lemma 2.7, the ξ-component of the MVSP (if it exists) must satisfy

$$\xi_t[0,x] = \tilde{\xi}(t,x) := \Gamma_1[\alpha[0,x] - \mu + \iota](t), \qquad t, x \in \mathbb{R}_+, \tag{2.16}$$

and the β-component must satisfy

$$\beta_t(x,\infty) = \tilde{\beta}(t,x) := \Gamma_2[\alpha[0,x] - \mu + \iota](t), \qquad t, x \in \mathbb{R}_+, \tag{2.17}$$

which, together with (2.10) of Lemma 2.7, shows that $\beta[0,\infty)$ must satisfy the relation

$$\beta_t[0,\infty) = \tilde{\beta}(t,0) + \alpha_t(\{0\}) - \tilde{\xi}(t,0). \tag{2.18}$$

Since (2.16) and (2.17) imply (2.8), and (2.17) and (2.18) imply (2.9), by Lemma 2.7, (ξ, β, ι) satisfy properties (1)-(4) of Definition 2.5. Thus, to show that (ξ, β, ι) solve the MVSP for (α, μ), it suffices to show that the quantities ξ and β defined via (2.16)–(2.18) lie in the right spaces: namely, that (a) for $t > 0$, $\xi_t \in \mathcal{M}$, $\beta_t \in \mathcal{M}$, and (b) $\xi \in \mathbb{D}_{\mathcal{M}}$ and $\beta \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$.
We start by establishing three assertions that clearly imply property (a).
(i) For $t \geq 0$, the map $x \mapsto \tilde{\xi}(t,x)$ lies in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$ and the map $x \mapsto \tilde{\beta}(t,x)$ is right-continuous, non-negative and non-increasing; moreover, both maps are continuous if $\alpha \in \mathcal{M}_0$;
(ii) the map $t \mapsto \alpha_t(\{0\}) - \tilde{\xi}(t,0)$ lies in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$;
(iii) $\sup_{t\in[0,T]} \sup_x \tilde{\xi}(t,x) < \infty$ and for $t \geq 0$,

$$\lim_{x\to\infty} \tilde{\xi}(t,x) = \Gamma_1[\alpha[0,\infty) - \mu + \iota](t) = \alpha_t[0,\infty) - \mu(t) + \iota(t). \tag{2.19}$$

To prove assertion (i), fix $t \geq 0$ and first notice that by (2.16), (2.17) and the definition of the SM Γ, $\tilde{\xi}(t,x)$ and $\tilde{\beta}(t,x)$ are non-negative for every $x \geq 0$. Next, we establish the monotonicity of $\tilde{\xi}(t,\cdot)$ and $\tilde{\beta}(t,\cdot)$. Let $0 \leq x_1 \leq x_2 < \infty$ and for $j = 1,2$, define $\psi_j := \alpha[0,x_j] - \mu + \iota$. Then $\psi_2 - \psi_1 = \alpha(x_1, x_2]$, which lies in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$ by Lemma 2.4(2). Therefore, by (2.16), (2.17) and the monotonicity property in Lemma 2.2(1), it follows that $\tilde{\xi}(t,x_2) - \tilde{\xi}(t,x_1) \geq 0$, and

$$t \mapsto \tilde{\xi}(t,x) \in \mathbb{D}_{\mathbb{R}}, \qquad t \mapsto \tilde{\beta}(t,x_1) - \tilde{\beta}(t,x_2) \in \mathbb{D}^{\uparrow}_{\mathbb{R}}, \qquad t \mapsto \tilde{\beta}(t,0) \in \mathbb{D}^{\uparrow}_{\mathbb{R}}. \tag{2.20}$$

The monotonicity property also shows that both $x \mapsto \tilde{\beta}(t,x)$ and $x \mapsto \tilde{\xi}(x,\cdot)$ have finite left limits on $(0,\infty)$. Next, to show right-continuity of $\tilde{\xi}(t,\cdot)$ and $\tilde{\beta}(t,\cdot)$, note that (2.16), (2.17) and the Lipschitz property in Lemma 2.2(2) imply

$$|\tilde{\xi}(t,x_2) - \tilde{\xi}(t,x_1)| \vee |\tilde{\beta}(t,x_1) - \tilde{\beta}(t,x_2)| \leq 2||\psi_2 - \psi_1||_t \leq 2 \sup_{s\in[0,t]} |\alpha_s(x_1,x_2]|. \tag{2.21}$$

Sending $x_2 \downarrow x_1$, the right-hand side goes to zero by Lemma 2.4(3) and the fact that $\alpha \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$. This completes the proof of the first assertion of (i). For the second assertion, first we claim that (2.16) implies that for every $x \geq 0$, $\xi[0,x) = \Gamma(\psi^x)$, where $\psi^x \doteq \alpha[0,x) - \mu + \iota$. Indeed, this can be seen by taking a sequence $y_n \uparrow x$, and setting $\psi_n := \alpha[0,y_n] - \mu + \iota$, and noting that $\psi_n \to \psi$ uniformly on compacts due to (2.4). However, if $\alpha \in \mathbb{D}^{\uparrow}_{\mathcal{M}_0}$ then $\psi^x = \alpha[0,x] - \mu + \iota$ and, hence, for every $t \geq 0$, (2.16) and (2.17) show that $\xi_t[0,x] = \xi_t[0,x)$ for every $x \geq 0$, and $\beta_t(\{x\}) = 0$ for every $x > 0$. Moreover, then (2.18) shows that we also have $\beta_t(\{0\}) = 0$. This proves $\xi_t, \beta_t \in \mathcal{M}_0$ for every $t \geq 0$ and thus concludes the proof of (i).

To prove property (ii), using (2.18), as well as (2.16) and (2.17) with $x = 0$, we obtain

$$\beta(\{0\}) = \alpha(\{0\}) - \xi(\{0\}) = \mu - \iota + \beta(0,\infty).$$

From (2.17) with $x = 0$, the fact that $\alpha(\{0\}) - \mu + \iota \in \mathbb{D}_{\mathbb{R}}$, and the definition of Γ_2 in (2.1), it follows that $\beta(0,\infty) \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$. Thus, to establish the claim it suffices to show that $\mu - \iota \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$. Since $\mu \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$, it follows that $(0,\mu) = \Gamma(-\mu)$. Now, set $\psi_1 = -\mu$ and $\psi_2 = \alpha[0,\infty) - \mu$, and observe that $\iota = \Gamma_2(\psi_2)$, $\psi_2 - \psi_1 = \alpha[0,\infty) \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ and $\Gamma_2(\psi_1) - \Gamma_2(\psi_2) = \mu - \iota$. The claim then follows from Lemma 2.2(1), and thus property (ii) is proved.

To prove property (iii), note that by (2.16) and the definition of Γ_1 in (2.1), for every $x, t \in \mathbb{R}_+$, $\tilde{\xi}(t,x) \leq \alpha_t[0,x] + 2||\mu - \iota||_t$. Thus,

$$\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \tilde{\xi}(t,x) \leq \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \alpha_t[0,x] + 2||\mu - \iota||_T = \alpha_T[0,\infty) + 2||\mu - \iota||_T < \infty,$$

where the last equality uses the monotonicity of α. Next, to show (2.19), send $x \to \infty$ in (2.16), use the Lipschitz continuity of Γ_1 and the fact that $\alpha[0,x] \to \alpha[0,\infty)$, to see that the first equality in (2.19) holds. The second equality follows because (2.15) and the definition of Γ_2 imply that $\alpha[0,\infty) - \mu + \iota \geq 0$, which in turn implies that Γ_1 leaves $\alpha[0,\infty) - \mu + \iota$ invariant. This completes the proof of property (a).

We now turn to the proof of property (b). For any $0 \leq x < y$, (2.20) and (2.17) show that $t \mapsto \beta_t(0,x]$ and $t \mapsto \beta_t(x,y]$ lie in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$, and from property (ii) above, (2.18) and (2.17) we see that $t \mapsto \beta_t(\{0\}) \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$. Thus, for every $x \geq 0$, $\beta.[0,x] \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$. To prove property (b), it suffices show that $\beta \in \mathbb{D}_{\mathcal{M}}$ because then Lemma 2.4(2) implies $\beta \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$ and, since (ξ, β, ι) satisfy properties 1 and 4 of Definition 2.5, (2.6) and the fact that $\alpha \in \mathbb{D}_{\mathcal{M}}$ imply $\xi \in \mathbb{D}_{\mathcal{M}}$. To show $\beta \in \mathbb{D}_{\mathcal{M}}$, fix a sequence $\{s_n\} \subset \mathbb{R}_+$. If $s_n \downarrow s$ for some $s \geq 0$, then for every $x \in [0,\infty)$, the fact that $\beta.[0,x] \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ implies $\beta_{s_n}[0,x] \to \beta_s[0,x]$, which proves that $\beta_{s_n} \to \beta_s$ in $\mathcal{M}$. We now show that $t \mapsto \beta_t \in \mathcal{M}$ has left limits. Next, fix $s > 0$ and a sequence $\{s_n\}$ such that $s_n \uparrow s$. For every $x \geq 0$, the fact that $\beta.[0,x] \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ implies that $\tilde{\nu}(x) \doteq \lim_{s_n \uparrow s} \beta_{s_n}[0,x]$ exists and is finite. It only remains to show that $x \mapsto \tilde{\nu}(x)$ lies in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$, since this would imply that $\beta_{s_n} \to \nu \in \mathcal{M}$, where $\nu[0,x] := \tilde{\nu}(x)$, $x \geq 0$. The monotonicity (and therefore existence of finite left limits) of $\tilde{\nu}$ follows immediately from the monotonicity of $x \mapsto \beta_t[0,x]$ for each $t \geq 0$. Also, given the monotonicity and right continuity of $t \mapsto \beta_t$ in $\mathcal{M}$ established above, it follows from (2.4) that for every $x \geq 0$, $\lim_{x_k \downarrow x} \sup_{n\in\mathbb{N}} \beta_{s_n}(x, x_k] = 0$, which in turn implies that $|\tilde{\nu}(x_k) - \tilde{\nu}(x)| \to 0$ as $x_k \downarrow x$. This completes the proof that $\tilde{\nu} \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ and establishes property (b) and hence, the first assertion of the proposition.

The second assertion follows from the first due to (2.7). The last assertion can be proved using arguments exactly analogous to those used in the proof of the first assertion (using the fact that the SM Γ maps $\mathbb{C}_{\mathbb{R}}$ into $\mathbb{C}_{\mathbb{R}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$), and is thus omitted. □

Given the uniqueness result of Proposition 2.8 we can now define the MVSM.

Definition 2.9 *(MVSM) Let $\Theta : \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}} \to \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ denote the map that takes $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ to the unique solution $(\xi, \beta, \iota) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ of the MVSP. We will refer to Θ as the MVSM.*

We now establish some regularity properties of the MVSM.

Proposition 2.10 *The map Θ satisfies the following two properties.*

1. *Suppose the sequence (α^k, μ^k), $k \in \mathbb{N}$, converges in $\mathbb{D}_{\mathcal{M}\times\mathbb{R}}$ to $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}_0} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$. Then $\Theta(\alpha^k, \mu^k) \to \Theta(\alpha, \mu)$ in $\mathbb{D}_{\mathcal{M}\times\mathcal{M}\times\mathbb{R}}$. In particular, Θ is continuous on $\mathbb{C}^{\uparrow}_{\mathcal{M}_0} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$.*

2. *The map $\Theta : \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}} \mapsto \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ is measurable.*

Proof: To prove the first property, let $(\xi^k, \beta^k, \iota^k) \doteq \Theta(\alpha^k, \mu^k)$, $k \in \mathbb{N}$, and let $(\xi, \beta, \iota) \doteq \Theta(\alpha, \mu)$. Then by Lemma 2.7, it follows that for every $x \geq 0$,

$$\iota = \Gamma_2[\alpha[0,\infty) - \mu] \qquad \text{and} \qquad (\xi[0,x], \beta(x,\infty) + \iota) = \Gamma[\alpha[0,x] - \mu], \tag{2.22}$$

and for every $k \in \mathbb{N}$,

$$\iota^k = \Gamma_2[\alpha^k[0,\infty) - \mu^k] \qquad \text{and} \qquad (\xi^k[0,x], \beta^k(x,\infty) + \iota^k) = \Gamma[\alpha^k[0,x] - \mu^k]. \tag{2.23}$$

Fix $0 < T < \infty$. Since $(\alpha^k, \mu^k) \to (\alpha, \mu)$ in $\mathbb{D}_{\mathcal{M}\times\mathbb{R}}$, there exists a strictly increasing continuous bijection $\tau^k : [0,T] \mapsto [0,T]$ with $\sup_{t\in[0,T]} |\tau^k(t) - t| \to 0$ such that

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \left[d_{\mathcal{L}}(\alpha^k_{\tau^k(t)}, \alpha_t) + |\mu^k(\tau^k(t)) - \mu(t)| \right] = 0.$$

Since we have $\alpha \in \mathbb{D}^{\uparrow}_{\mathcal{M}_0}$, it follows that for every $\varepsilon > 0$, $\sup_{t\in[0,T]} Osc_\varepsilon(\alpha_t[0,\cdot]) \leq Osc_\varepsilon(\alpha_T[0,\cdot])$ and $\lim_{\varepsilon\downarrow 0} Osc_\varepsilon(\alpha_T[0,\cdot]) = 0$. Therefore, combining the last display with the inequality (1.2), we obtain

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \sup_{x\in[0,\infty)} \left| \alpha^k_{\tau^k(t)}[0,x] - \mu^k(\tau^k(t)) - (\alpha_t[0,x] - \mu(t)) \right| = 0.$$

Together with (2.22), (2.23), the fact that $\varphi = \Gamma(\psi)$ implies $\varphi \circ \tau^k = \Gamma(\psi \circ \tau^k)$ and the Lipschitz continuity of Γ from Lemma 2.2(2), this implies

$$\lim_{k\to\infty} \sup_{t\in[0,T]} |\iota(\tau^k(t)) - \iota(t)| = 0, \tag{2.24}$$

and

$$\lim_{k\to\infty} \sup_{t\in[0,T]} \sup_{x\in[0,\infty)} \max \Big(\left| \xi^k_{\tau^k(t)}[0,x] - \xi_t[0,x] \right|, \\ \left| \beta^k_{\tau^k(t)}(x,\infty) - \iota(\tau^k(t)) - (\beta_{\tau(t)}(x,\infty) - \iota(t)) \right| \Big) = 0. \tag{2.25}$$

From (2.24) we have $\iota^k \to \iota$ in $\mathbb{D}_{\mathbb{R}}$, and from (2.25) and (1.2) it follows that one also has $\sup_{t\in[0,T]} d_{\mathcal{L}}(\xi^k_{\tau^k(t)}, \xi_t) \to 0$. Since $\sup_{t\in[0,T]} |\tau^k(t) - t| \to 0$, by the definition of the Skorokhod topology, it follows that $\xi^k \to \xi$ in $\mathbb{D}_{\mathcal{M}}$. Since, by Proposition 2.8, $\alpha \in \mathbb{D}_{\mathcal{M}_0}$ implies $\xi, \beta \in \mathbb{D}_{\mathcal{M}_0}$, it follows that $\xi(\{0\}) = \beta(\{0\}) = 0$ and hence, $\xi^k(\{0\}) \to 0$. The fact that Lemma 2.7 implies that (2.9) holds with α, β, ξ replaced by α^k, β^k, ξ^k, respectively, then implies that $\beta^k(\{0\}) \to 0$, which together with (2.25) and (2.24), implies $\beta^k \to \beta$ in $\mathbb{D}_{\mathcal{M}}$. This proves the first assertion of the first property. The second assertion is an immediate consequence of the first and the fact that if $(\alpha^k, \mu^k) \to (\alpha, \mu)$ in the product topology $\mathbb{D}_{\mathcal{M}} \times \mathbb{D}_{\mathbb{R}}$, and $(\alpha, \mu) \in \mathbb{C}_{\mathcal{M}} \times \mathbb{C}_{\mathbb{R}}$, then $(\alpha^k, \mu^k) \to (\alpha, \mu)$ in $\mathbb{D}_{\mathcal{M}\times\mathbb{R}}$.

We now turn to the proof of the second property, namely the measurability of Θ. It is clearly enough to establish the measurability of each component map. The proof for the third component is easy. Since properties 2 and 4 of the MVSP imply that $\iota = \Gamma_2(\alpha[0,\infty) - \mu)$, Γ_2 is continuous, the maps $\alpha \mapsto \alpha[0,\infty)$ and $(\alpha[0,\infty), \mu) \mapsto \alpha[0,\infty) - \mu$ are measurable, it follows that ι is a measurable function of (α, μ). Moreover, in view of Remark 2.6, specifically the balance equation (2.6), and the fact that addition map from $\mathbb{D}_{\mathcal{M}} \times \mathbb{D}_{\mathcal{M}} \mapsto \mathbb{D}_{\mathcal{M}}$ is measurable, measurability of the second component follows from that of the first. To show measurability of the first component, by Lemma 2.4(4), we only need to show that for every $t, x \geq 0$, the map $\mathcal{T} : \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}} \to \mathcal{M}$, defined by $\mathcal{T}(\alpha, \mu) = \xi_t[0, x]$ is measurable. But this follows from (2.22), the measurability of the maps $(\alpha, \mu) \mapsto \alpha[0, x] - \mu$, $x \geq 0$, and the continuity of Γ_1. □

Remark 2.11 It is worthwhile to contrast the MVSM with another Skorokhod-type map that was introduced in [1], which considered a generalization of the SM in which the time interval $[0, \infty)$ is replaced by a general poset (partially ordered set), and a function on the poset is constrained in a minimal fashion to lie within two prescribed functions on the poset. In the special case when the poset is $\mathbb{R}_+$ and the prescribed functions are constant functions with values $a < b$, this reduces to the Skorokhod map on $[a, b]$, also referred to as the double-barrier Skorokhod map. When instead, the poset is chosen to be $\mathbb{T} \doteq [0, \infty) \times \mathcal{B}(\mathbb{R}_+)$, with the natural partial ordering $(t, A) \prec (\tilde{t}, \tilde{A})$ if and only if $t \leq \tilde{t}$ and $A \subseteq \tilde{A}$, then the map in [1] yields a map on measure-valued paths. Specifically, the pair $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ can be identified with the function $(t, A) \mapsto \alpha_t(A) - \mu(t)$ on the poset $\mathbb{T}$. However, the image of this function under the map of [1] with $a = 0$ and $b = \infty$, will correspond to (α, μ), providing, roughly speaking, a Jordan decomposition of the signed measure $\alpha - \mu$. This does not capture the dynamics we are interested in and obtain from the MVSP, where, in particular, the temporal component and the space component play different roles.

3 Some Illustrative Examples

In this section, we describe some simple examples that motivate the form of the MVSM that was introduced in Section 2.2. This section can be skipped without loss of continuity. We start in Section 3.1 by describing the K-class model with priorities and show how it can be characterized by K coupled SMs on the half-line, and in Section 3.2 we show how the MVSM arises naturally when trying to characterize a continuum version of the K-class model. In Section 3.3, we briefly show how two additional policies, First-In-First-Out (FIFO) and Last-In-First-Out (LIFO), can be expressed in terms of the MVSM . The discussion in this section is purely formal, and simply serves to emphasize that the MVSM and its relatives arise naturally as a tool for the analysis of queueing models with (continuum) priorities, and thus are likely to be useful beyond the specific models, EDF, SJF and SRPT, that are considered in detail Sections 4 and 5 of this paper.

3.1 The K-class Fluid Model With Priorities

Consider a queueing system that consists of K classes of jobs, each with a dedicated buffer that is fed by an external fluid arrival stream, and a single common server that can process material from the buffers at some specified (maximal) rate $\mu(t) \geq 0$ at time t. Let $x_i \geq 0$, $i = 1, \ldots, K$, represent the inital content of the class i buffer, and let $X_i(t)$ denote the (non-negative) content of the class i buffer at time $t \geq 0$. Let $\hat{A}_i$ be a non-decreasing function such that $\hat{A}_i(t)$ represents the total cumulative mass that arrived into buffer i during the time interval $[0, t]$. Assume that the priorities are ordered such that each class i has priority over all classes $j > i$. This means that the server can remove content from a class j buffer only when all class i buffers, $i < j$, are empty. The functions $A = (A_i)$, $A_i(\cdot) := x_i + \hat{A}_i(\cdot)$ and $M(\cdot) := \int_0^{\cdot} \mu(s)ds$ are regarded to be the problem data for this model, which we will call the *K-class model with priorities.*

For this model, it is possible to write down a set of equations and conditions that uniquely characterize $X = (X_i)$ in terms of the problem data (A, M). To this end, we now introduce some basic notation. Recall from Section 1.1 that $\mathbb{R}_+ = [0, \infty)$, $\mathbb{D}_{\mathbb{R}} = \mathbb{D}_{\mathbb{R}}(\mathbb{R}_+)$ is the space of functions from $\mathbb{R}_+$ to $\mathbb{R}$ that are right continuous with finite left limits on $(0, \infty)$, endowed with the Skorokhod J_1 topology, and $\mathbb{D}_{\mathbb{R}}^{\uparrow}$ is the subspace of non-decreasing functions in $\mathbb{D}_{\mathbb{R}}$. For data $(A, M) \in (\mathbb{D}_{\mathbb{R}}^{\uparrow})^{K+1}$ the model is fully described by the following three relations:
(i) *the balance equation between arrivals and departures*: there exist $B_i, 1 \le i \le K$, such that

$$
\begin{aligned}
&X_i(t) = A_i(t) - \int_{[0,t]} B_i(s)dM_s \ge 0, \quad \text{for } 1 \le i \le K, \\
&\text{where} \\
&0 \le B_i(t) \le \mathbb{I}_{\{X_i(t)>0\}}, \quad \sum_{i=1}^K B_i(t) \le 1.
\end{aligned}
\tag{3.1}
$$

Here, $B_i(t)$ represents the fraction of the server's effort that is dedicated to class i at time t.
(ii) a standard *work conservation* condition, which ensures that the server works at maximal capacity whenever there is content in any buffer:

$$
\sum_{i=1}^K X_i(t) > 0 \text{ implies } \sum_{i=1}^K B_i(t) = 1, \text{ for } dM\text{-a.e. } t \in [0, \infty); \tag{3.2}
$$

(iii) the *priority* condition:

$$
\text{for } 1 \le i < j \le K,\ X_i(t) > 0 \text{ implies } B_j(t) = 0, \text{ for } dM\text{-a.e. } t \in [0, \infty). \tag{3.3}
$$

For convenience, we also define the idleness process I as follows:

$$
I := 1 - \sum_{i=1}^K B_i. \tag{3.4}
$$

We now show that one can solve for $X \in \mathbb{D}_{\mathbb{R}}^K$ using repeated applications of the SM on the half-line defined in Section 2.1. First, for $H = X, A, B$, and $1 \le i < j \le K$, denote $H[i, j] = \sum_{k=i}^j H_k$, and set

$$
\begin{aligned}
\hat{M}_i(\cdot) &:= \int_{[0,\cdot]} B[i+1, K](s)dM_s, \qquad i = 0, \dots, K-1, \\
\hat{I}(\cdot) &:= \int_{[0,\cdot]} I(s)dM_s.
\end{aligned}
$$

Then, equations (3.1)–(3.3), imply that $\hat{M}_i$, $i = 0, \dots, K-1$, and $\hat{I}$ are all members of $\mathbb{D}^{\uparrow}_{\mathbb{R}}(\mathbb{R}_+)$, and, with $\hat{M}_K := 0$,

$$\begin{array}{rcll} X[1,i] &=& A[1,i] - M + \hat{M}_i + \hat{I} \;\geq\; 0, & i = 1, \dots, K, \\ \hat{M}_0 + \hat{I} &=& M, & \\ \text{where} && & \\ X[1,i] &=& 0 \; \mathrm{d}\hat{M}_i\text{-a.e.}, & i = 1, \dots, K-1, \\ X[1,i] &=& 0 \; \mathrm{d}\hat{I}\text{-a.e.}, & i = 1, \dots, K. \end{array} \tag{3.5}$$

Comparing this set of equations with the SP on the half-line from Definition 2.1, it clearly follows that

$$(X[1,i], \hat{M}_i + \hat{I}) = \mathit{\Gamma}_1\big[A[1,i] - M\big], \qquad i = 1, \dots, K,$$

which is exactly analogous to (2.8). Thus, we have shown that the buffer content process for the fluid queue with a finite number of priority classes can be "solved" using a finite number of applications of the SM on the half-line.

3.2 The continuum-priority fluid queue

We now consider the formal limit of the K-class model with priorities, as K, the number of classes, increases to infinity and the arrival rate to each class is scaled down by a factor $1/K$. With a view to describing such a limit, first, for each finite K, note that we can map the set of classes in the K-class model to the interval $[0,1]$ by identifying each class $i \in \{1, \dots, K\}$ with the number $1/i \in \{1/K, \dots, 1\} \subset [0,1]$. The priority rule then translates to the condition that for each $x \in (0,1]$, any class within $[0,x]$ has priority over every class within $(x,1]$. In the continuum limit model, priority classes are indexed by $[0,1]$ and the above priority rule continues to hold. Moreover, we assume that arrivals are governed by some measurable, locally integrable function $\lambda : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$, where $\lambda(t,x)dtdx$ can be regarded as the quantity of arrivals during the time interval $[t, t+dt]$, into classes within the interval $[x, x+dx]$. Then, we can define the cumulative arrival stream for the fluid model, α, to be

$$\alpha_t[0,x] = \int_{[0,t]\times[0,x]} \lambda(s,y)dsdy, \qquad t \in \mathbb{R}_+, x \in [0,1].$$

Setting $\alpha_t[0,x] = \alpha_t[0,1]$ for all $x > 1$, we obtain a well-defined path $\alpha \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$. We also assume, as before, that we are also given a function $\mu \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$, where $\mu(t)$ represents the maximal amount of mass a server could process in the interval $[0,t]$.

We now show that, just as a finite number of coupled SMs on the half-line were useful for describing the solution to the K-class priority model, the limiting continuum priority fluid model is naturally described by the MVSM . For $t \geq 0$, let ξ_t and β_t be measures on $\mathbb{R}_+$, where $\xi_t[x, x+dx]$ denotes the quantity of jobs with priority $[x, x+dx]$ that at time t are in the queue, and $\beta_t[x, x+dx]$ represents the quantity of jobs from classes in $[x, x+dx]$ that have been served by time t and let $\iota(t)$ be a real-valued function that represents the cumulative idleness time of the server in the interval $[0,t]$. Comparing the description of the continuum-priority model with Definition 2.9 of the MVSM $\mathit{\Theta}$, it is not hard to arrive at the following fluid model equation for the continuum priority model:

$$(\xi, \beta, \iota) = \mathit{\Theta}(\alpha, \mu). \tag{3.6}$$

Thus, in this case, the fluid model is fully described by specifying the data (α, μ) and considering equation (3.6).

3.3 FIFO and LIFO

We now briefly introduce two other well-known single-server queueing models that can also be described in terms of the MVSM and its close relatives. Here, we assume we are given a measurable function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$, where $\lambda(t)dt$ represents the arrivals during the time interval $(t, t+dt)$, and the server prioritizes jobs in the queue in the order of their arrival (FIFO) or in reverse order (LIFO). We thus let

$$\alpha_t[0,x] = \int_{[0,t\wedge x]} \lambda(s)ds, \quad t \in \mathbb{R}_+,\, x \in \mathbb{R}_+.$$

For the FIFO discipline, the same logic as earlier then yields the equation

$$(\xi, \beta, \iota) = \Theta(\alpha, \mu). \tag{3.7}$$

For the case of LIFO, one has to redefine Θ by performing inversion with respect to the x variable. Specifically, suppose we consider a modified version of Definition 2.5, in which items 3 and 4 are the same as before, but items 1 and 2 are modified as follows: for $x > 0$,
1'. $\xi(x,\infty) = \alpha(x,\infty) - \mu + \beta[0,x] + \iota$,
2'. $\xi(x,\infty) = 0 \quad d\beta[0,x)$-*a.e.*
Analogous to the MVSM , it can be shown that there exists a unique map Θ' that satisfies items 1', 2', 3 and 4 and the LIFO model dynamics would then be captured by the equation (3.7), but with Θ replaced by Θ'.

4 Fluid models

In this section, we present fluid models of three classes of queueing models in which service is prioritized according to a continuous parameter. In the case of the EDF policy, which is considered in Section 4.1, the continuous parameter is the job's deadline, while for the SJF and SRPT policies considered in Section 4.2, it is the remaining processing time. In each case, we include some heuristic discussion to provide intuition into the form of the fluid model equations and show that it can be represented in terms of the MVSM Θ; rigorous convergence of a sequence of scaled stochastic models to the fluid model is established in Section 5.1 (see Theorem 5.4) for EDF and Section 5.2 (see Theorems 5.13 and 5.16) for SJF and SRPT.

4.1 Earliest-Deadline-First Fluid Model

Section 4.1.1 introduces the state descriptors of the non-preemptive hard EDF fluid model described in the introduction, and the associated fluid model equations (the corresponding stochastic model is described in Section 5.1.1). Section 4.1.2 and Section 4.1.3 provide two alternative formulations of the fluid model equations, which are shown to be equivalent in Section 4.1.4 under additional assumptions on the data. Section 4.1.5 provides an optimality result of the fluid model EDF in terms of the reneging count.

4.1.1 Description of the EDF Fluid Model

We now consider the non-preemptive soft and hard EDF models described in the introduction, in which jobs arrive at a buffer that has infinite room and on arrival, declare their deadlines, which represents the time by which the job should enter service. In addition, jobs may be present initially, that is at time zero, and their deadlines are assumed to be known. The server can serve at most one job at a time and, when it becomes available, chooses in a non pre-emptive fashion to serve the job with the least deadline among those that are still in the system. (Ties may be assumed to be broken by giving priority to the job with the earlier arrival time, although the details of this mechanism are not relevant for the fluid model.) In particular, the server never idles when there are jobs in the system. In the soft EDF model, jobs wait to be served even after their deadline has elapsed, whereas in the hard EDF model, a job that does not start service prior to its deadline leaves the system. We will use the term *departure* to refer to jobs that leave the system on completion of service and the term *reneging* to refer to jobs that exit the system on reaching their deadline without starting service. Jobs do not renege while being served. In a fluid model, given a Borel set $A \subset \mathbb{R}_+$, we let $\hat{\alpha}_t(A)$ denote the mass of jobs that have arrived up to time t with deadlines in the set A. It is worth emphasizing that here, we consider *absolute* deadlines, as opposed to some other works (e.g. [9, 26]), which consider *relative* deadlines, also referred to as lead times, which are defined as the difference between the deadline and the current time. In other words, in our system the deadline of a job does not change with time and, under the hard EDF policy, a job with deadline x reneges at time x if it did not enter service earlier; this is in contrast to relative deadlines, which decrease with time, and if a job has a relative deadline x at time t, then it would renege at time $t+x$ if it does not enter service before that time. Note that the absolute deadline of a job coincides with its relative deadline only at the time the job arrives to the system. Here, and in what follows, 'deadline' will be used to mean 'absolute deadline'. The term *patience* will be used to mean the time that a job is willing to wait when it arrives; that is, the (absolute) deadline is equal to the time of arrival plus the patience. It is common to assume that a job's patience follows a fixed distribution. In this case, the fluid arrival stream $\hat{\alpha}$ has a specific form; see (4.3) of Assumption 4.5. However, in this section we allow $\hat{\alpha}$ to be a generic member of $\mathbb{D}^{\uparrow}_{\mathcal{M}}$. We also let $\xi_{0-} \in \mathcal{M}$ represent the empirical distribution of deadlines corresponding to jobs that arrived before time 0 and are still in the system at time 0, and let $\alpha := \xi_{0-} + \hat{\alpha}$. To complete the specification of the model data, we assume that $\mu \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$, where $\mu(t)$ represents the mass the server can potentially process in time $[0,t]$. We will refer to (α, μ) as the data for the fluid model.

We now introduce the quantities that describe the fluid model for this system. Given a measurable set A in $\mathbb{R}_+$ and $t \geq 0$, let $\xi_t(A)$ represent the mass of jobs in the buffer at time t that have deadline in the set A, and let $\iota(t)$ represent the total amount of unused potential service in $[0,t]$ due to server idleness. The quantity β has a slightly different interpretation. Specifically, the quantity $\beta_t(A)$ represents the mass of jobs with deadlines in A that by time t have left the queue: either by transferring to the server or (in the hard model) by reneging. In analogy with the continuum priority model described in Section 3.2, the state process is then (ξ, β, ι) and thus, the soft EDF fluid model is then concisely described by the equation (3.6). On the other hand, to fully describe the state of the hard EDF fluid model we need to introduce one additional function, $\rho \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$. For $t > 0$, the quantity $\rho(t)$ represents the total amount of mass that has left the system by reneging in the interval $[0,t]$. The (*a priori*) unknown system state descriptor or fluid model solution for the hard EDF policy is then represented by $(\xi, \beta, \iota, \rho)$.

From the description of the policy and the definition of the MVSM Θ, it is reasonable to expect that the state $(\xi, \beta, \iota, \rho)$ should satisfy the following set of equations:

$$\begin{cases} (i) & (\xi, \beta, \iota) = \Theta(\alpha, \mu + \rho), \\ (ii) & \xi_t[0,t) = 0, \quad \text{for every } t > 0, \end{cases} \tag{4.1}$$

where property (ii) captures the condition that any job with a deadline strictly less than t would have been served or would have reneged from the system by time t. However, these equations are not sufficient to uniquely characterize the model; in particular, they put no constraints on ρ. We now identify two additional conditions that we would expect ρ to satisfy given the description of the policy. The first one is a minimality condition, described in Section 4.1.2, and shown to be satisfied by the hard EDF policy in Theorem 4.10. In particular, Theorem 4.10 establishes an optimality result for the (hard) EDF fluid model, showing that it leads to the least amount of reneged work in the system amongst a reasonably large class of policies (a precise statement appears in Section 4.1.5). The second condition, introduced in Section 4.1.3, imposes the requirement that ρ increases only on the set of times t at which the left end of the support of ξ_t equals t. This captures the property that, under the hard EDF policy described above, if a job reneges, it does so exactly at the time of its deadline. In Section 4.1.4 we show that, under natural additional assumptions on the data, the two formulations are equivalent.

4.1.2 A Minimal Solution

We introduce the notion of a minimal solution of (4.1), and show that it is well defined.

Definition 4.1 (Minimal Solution) *A solution $(\xi, \beta, \iota, \rho)$ of (4.1) is said to be minimal if for every solution $(\xi^1, \beta^1, \iota^1, \rho^1)$ of (4.1), one has $\rho \leq \rho^1$, that is, $\rho(t) \leq \rho^1(t)$ for every $t \geq 0$.*

Proposition 4.2 *Given $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$, there exists a unique minimal solution $(\xi, \beta, \iota, \rho) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ of (4.1).*

Proof: Uniqueness is an immediate consequence of minimality: if ρ^1 and ρ^2 are two minimal solutions then they must satisfy $\rho^1 \leq \rho^2 \leq \rho^1$ and hence, they must be equal.

Next, we construct a minimal solution in the form of the lower envelope of the collection of all solutions. Fix $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$. Let $\mathbb{S}$ denote the collection of all $\rho \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ for which there exists $(\xi, \beta, \iota) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ such that $(\xi, \beta, \iota, \rho)$ is a solution of (4.1). First, note that $\mathbb{S}$ is nonempty. Indeed, let $\rho_t = \alpha_t[0,\infty)$. Then, since $\xi = \Theta(\alpha, \mu + \rho)$, by (2.10), $\xi[0,\infty) = \Gamma_1[\alpha[0,\infty) - \mu - \rho] = \Gamma_1[-\mu] = 0$ where we used the fact that $-\mu$ is decreasing and nonpositive and Lemma 2.3(2). Thus, $\xi \equiv 0$, and so (4.1)(ii) is automatically satisfied.

Now, for $t \in [0,\infty)$, let $\bar{\rho}(t) := \inf\{\rho(t) : \rho \in \mathbb{S}\}$. It is not hard to verify that the infimum of a collection of non-negative, non-decreasing and right-continuous functions also possesses the same properties. Indeed, this can be verified directly or deduced from the fact that a non-decreasing function with left limits is right-continuous if and only if it is upper semicontinuous, and the infimum of upper semicontinuous functions (resp. non-decreasing) is upper semicontinuous (resp. non-decreasing). Thus, we have shown that $\bar{\rho} \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$.

By definition, for every $\rho \in \mathbb{S}$, we have $\bar{\rho}(t) \leq \rho(t)$ for all $t \geq 0$. Now, set $(\bar{\xi}, \bar{\beta}, \bar{\iota}) := \Theta(\alpha, \mu + \bar{\rho})$. Then, to show that $(\bar{\xi}, \bar{\beta}, \bar{\iota}, \bar{\rho})$ is a minimal solution of (4.1), it only remains to prove that $\bar{\xi}_t[0,t) = 0$ for every $t > 0$. Fix $t > 0$ and $x \in (0,t)$. Let $(\xi, \beta, \iota, \rho)$ be a solution to (4.1). Then, by Lemma 2.7 we have $\xi_t[0,x] = \Gamma_1[v - \rho](t)$, where for notational convenience we set $v(s) := \alpha_s[0,x] - \mu(s)$ for $s \geq 0$. Moreover, since $x < t$, by (4.1)(ii) we have $\xi_t[0,x] = 0$. In turn, by the explicit form of Γ_1 given in (2.1) it follows that $v(t) - \rho(t) = \inf_{s \in [0,t]}(v(s) - \rho(s))$. Hence, we have for $s \in [0,t]$,

$$v(t) - \bar{\rho}(t) = \sup_{\rho \in \mathbb{S}}(v(t) - \rho(t)) \leq \sup_{\rho \in \mathbb{S}}(v(s) - \rho(s)) = v(s) - \bar{\rho}(s).$$

This implies that $v(t) - \bar{\rho}(t) = \inf_{s \in [0,t]}(v(s) - \bar{\rho}(s)) \wedge 0$, and so the definition of Γ_1 in (2.1) shows that $\Gamma(v - \bar{\rho})(t) = 0$. Since this holds for every $x \in (0,t)$, $\bar{\xi}_t[0,t) = 0$. This completes the proof that $(\bar{\xi}, \bar{\beta}, \bar{\iota}, \bar{\rho})$ is a minimal solution. □

As a first application of Proposition 4.2, we obtain an intuitive monotonicity property of the reneging count ρ with respect to the data (α, μ). It is closely related to a result obtained in [27] for the $G/M/1+G$ queue in the setting of a stochastic recursive sequence. Roughly speaking, it states that reneging is monotoncally increasing [resp., decreasing] w.r.t. the cumulative arrival [resp., service function]. The ordering in [27] is obtained with respect to the patience time distribution function.

Corollary 4.3 *Let $(\alpha^i, \mu^i) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}_+}, i = 1,2$ be such that $(\alpha^1_t[0,x] - \mu^1_t) - (\alpha^2_t[0,x] - \mu^2_t)$ is non-negative and non-decreasing in t for every $x \in \mathbb{R}_+$. Denote by $(\xi^i, \beta^i, \iota^i, \rho^i)$ the unique minimal solution of* (4.1) *corresponding to $(\alpha^i, \mu^i), i = 1,2$. Then we have $\rho^1 \geq \rho^2$.*

Proof: Let $(\tilde{\xi}, \tilde{\beta}, \tilde{\iota}) = \Theta(\alpha^2, \mu^2 + \rho^1)$. Now, for every x, $\xi^1[0,x] = \Gamma_1[\alpha^1[0,x] - \mu^1 - \rho^1]$, while $\tilde{\xi}[0,x] = \Gamma_1[\alpha^2[0,x] - \mu^2 - \rho^1]$, and therefore using Lemma 2.2(1), $\tilde{\xi} \leq \xi^1$. As a result, $\tilde{\xi}_t[0,t) = 0$ must hold for all $t > 0$. This shows that $(\tilde{\xi}, \tilde{\beta}, \tilde{\iota}, \rho_1)$ is a solution of (4.1) corresponding to (α^2, μ^2). Thus by minimality of $(\xi^2, \beta^2, \iota^2, \rho^2)$, we obtain $\rho^1 \geq \rho^2$. □

4.1.3 Hard EDF Fluid Model Equations

We now present the fluid model equations for the hard EDF policy.

$$\begin{cases} (i) & (\xi, \beta, \iota) = \Theta(\alpha, \mu + \rho), \\ (ii) & \xi_t[0,t) = 0, \quad \text{for every } t > 0, \\ (iii) & \sigma(t) = t \quad d\rho\text{-a.e., where for } t \geq 0,\ \sigma(t) = \min \operatorname{supp}[\xi_t]. \end{cases} \tag{4.2}$$

For property (4.2)(iii) to be well defined, σ needs to be a measurable function. The next lemma establishes this property.

Lemma 4.4 *Given $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$, suppose $(\xi, \beta, \iota) = \Theta(\alpha, \mu + \rho)$ for some $\rho \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$. Then, for every $t \geq 0$, the map $a \mapsto \xi_t[0, t+a]$ from $[0,\infty)$ to $[0,\infty)$ is right continuous, and for every $a \geq 0$, the map $t \mapsto \xi_t[0,t+a]$ is right continuous. Moreover, if $\xi_t[0,t) = 0$ and $\sigma(t) = \min \operatorname{supp}[\xi_t]$, $t \geq 0$, then the mapping $\sigma : [0,\infty) \mapsto \mathbb{R} \cup \{\infty\}$ is measurable.*

Proof: For fixed $t \in [0, \infty)$, the right continuity of $a \mapsto \xi_t[0, t+a]$ follows from the fact that ξ_t is a finite measure. For fixed $a \in [0, \infty)$, to show the right continuity of $t \mapsto \xi_t[0, t+a]$, fix any sequence $\{t_n\}$ in $[0, \infty)$ such that $t_n \downarrow t$. Then, by Lemma 2.7, $\xi[0, x] = \Gamma_1[\alpha[0, x] - \mu - \rho]$, the explicit expression for Γ_1 in (2.1), and the fact that $\alpha[0, x], \mu, \rho$ are non-decreasing, we have for $n \in \mathbb{N}$,

$$\begin{aligned}
&|\xi_{t_n}[0, t_n + a] - \xi_t[0, t + a]| \\
&\qquad = |\xi_{t_n}[0, t_n + a] - \xi_t[0, t_n + a]| + |\xi_t[0, t_n + a] - \xi_t[0, t + a]| \\
&\qquad \le \alpha_{t_n}[0, \infty) - \alpha_t[0, \infty) + \mu(t_n) + \rho(t_n) - \mu(t) - \rho(t) + |\xi_t[0, t_n + a] - \xi_t[0, t + a]|.
\end{aligned}$$

Sending $n \to \infty$, the right-hand side goes to zero because the functions $\alpha[0, \infty)$, μ and ρ are right continuous, and ξ_t is a measure. This shows that $t \mapsto \xi_t[0, t + a]$ is right continuous. In turn, this right continuity together with the relations

$$\{t : \sigma(t) < t + u\} = \{t : \xi_t[0, t + u] > 0\} \text{ and } \{t : \sigma(t) = t\} = \bigcap_n \{t : \xi_t[0, t + n^{-1}] > 0\},$$

where the latter equality holds because $\xi_t[0, t) = 0$, implies the measurability of $t \mapsto \sigma(t)$. $\square$

We now show that under mild additional assumptions on the data (α, μ), the fluid model equations (4.2) have a unique solution that coincides with the minimal solution of (4.1).

Assumption 4.5 *Suppose the following two properties hold:*
(i) $\alpha = \hat{\alpha} + \xi_{0-}$*, where* $\xi_{0-} \in \mathcal{M}_0$*, and* $\hat{\alpha} \in \mathbb{C}^{\uparrow}_{\mathcal{M}_0}$ *satisfies*

$$\hat{\alpha}_t[0, x] = \int_0^t \mathbb{I}_{\{x \ge s\}} \nu_s[0, x - s] ds, \quad t \ge 0, x \ge 0, \tag{4.3}$$

for some measurable collection $\{\nu_s\}$ *of finite measures on* $\mathbb{R}_+$ *satisfying for every* $t \ge 0$ $\lim_{x \downarrow 0} \sup_{s \in [0,t]} \nu_s[0, x] = 0$*;*
(ii) there exist $\mu^0 \in \mathbb{C}^{\uparrow}_{\mathbb{R}}$ *and a non-negative measurable function* m *on* $[0, \infty)$ *satisfying* $\inf_{s \in [0,t]} m(s) > 0$ *for every* $t \in [0, \infty)$ *(i.e.,* m *is locally bounded away from zero), such that*

$$\mu(t) = \mu^0(t) + \int_0^t m(s) ds, \quad t \ge 0. \tag{4.4}$$

Remark 4.6 As mentioned earlier, the notation ξ_{0-} in Assumption 4.5 represents the state of the queue just prior to zero. The notation ξ_{0-} is used to emphasize that it need not coincide with ξ_0, which represents the state of the queue at time zero. In particular, the measures ξ_0 and ξ_{0-} may differ when μ has a jump at time zero, that is, when $\mu(0) > 0$.

Remark 4.7 When Assumption 4.5 holds, we will say that the data (α, μ) is associated with the primitives $(\xi_{0-}, \{\nu_s\}_{s \ge 0}, \mu^0, m)$. It is immediate from the expressions (4.3) and (4.4) and the stated properties of the primitives that (α, μ) lies in $(\mathbb{C}^{\uparrow}_{\mathcal{M}_0}, \mathbb{C}^{\uparrow}_{\mathbb{R}})$.

Remark 4.8 **(a)** To provide intuition into the assumed form (4.3) of the arrival process, note that $\hat{\alpha}_t[0, x]$ denotes the total amount of fluid that has entered the system by time t with absolute deadline in $[0, x]$. The indicator inside the integral in (4.3) captures the property that $\hat{\alpha}_t[0, x]$ cannot contain any fluid that has arrived after time x, since it would have absolute deadline more than x. The measure ν_s can be thought of as the *scaled* patience distribution of jobs that arrive at time s, where the *scale* factor represents the arrival rate at that time. The fact that any job arriving at time s has absolute deadline in $[0, x]$ if and only if its patience lies in $[0, x - s]$ explains the presence of the $\nu_s[0, x - s]$ term in the integral.

In particular, one can always write ν_s in (4.3) as $\lambda(s)\hat{\nu}_s$, where λ is scalar valued while each $\hat{\nu}_s$ is a probability measure. In this representation, $\lambda(s)$ corresponds to the rate of arrivals, while $\hat{\nu}_s$ gives the distribution of patience of jobs arriving at time s, as a function of s, that, naturally, is supported on $\mathbb{R}_+$. Note that our assumption on this collection of measures requires that its cdf is *uniformly* right-continuous at zero. A special case is when $\lambda(s)$ is generic while $\hat{\nu}_s = \hat{\nu}$ is fixed, corresponding to a fixed patience distribution.
(b) As a special case of Corollary 4.3, that fits the structure of the monotonicity result from [27], if α^i admits a form as in (4.3) with $\nu^i_s = \lambda(s)\nu^i, i = 1, 2$, and $\nu^1[0, x] \geq \nu^2[0, x]$ for all $x \in \mathbb{R}_+$, and $\mu^2 - \mu^1$ is non-decreasing with $\mu^2(0) - \mu^1(0) \geq 0$, then we have $\rho^1 \geq \rho^2$.
(c) In Section 5.1.2 we give further examples of data for the N-system that lead in the limit to the above form (4.3).

The condition $\inf_{s\in[0,t]} m(s) > 0$ in Assumption 4.5(ii) ensures that the system is always capable of processing fluid at a strictly positive rate. On the other hand, since $\hat{\alpha}_t[0, t] = 0$ and the arrival rate of fluid with deadline in $[0, x]$ is $\nu_s[0, x - s] \leq \nu_s[0, x]$, the condition $\lim_{x\downarrow 0} \sup_{s\in[0,t]} \nu_s[0, x] = 0$ in Assumption 4.5(i) ensures that the system is always capable of removing all the fluid with sufficiently small deadline that arrives into the system. The following lemma supplies a quantitative version of this statement.

Lemma 4.9 *If (α, μ) satisfy Assumption 4.5 then for any $\tau' < \infty$, there exists $\delta_0 \in (0, 1)$ such that for any $x \in [0, \delta_0]$ and $t'_0 \in [\tau', \tau' + \delta_0]$, the function $t \mapsto \alpha_t[0, t'_0 + x] - \mu(t)$ is non-increasing on $[\tau', \tau' + 2]$.*

Proof: Given any $\tau' < \infty$, Assumption 4.5(ii) implies that $c_0 := \inf_{u\in[0,\tau'+2]} m(u)$ is strictly positive. Assumption 4.5(i) then implies that there exists $\delta_0 \in (0, 1)$ sufficiently small so that $\sup_{u\in[0,\tau'+2]} \nu_u[0, 2\delta_0] < c_0$. Combining this with the expressions in (4.3) and (4.4) we then see that for any $t \geq 0$ and $x \in [0, \delta_0]$,

$$\begin{aligned} &\alpha_{\tau'+t}[0, t'_0 + x] - \alpha_{\tau'}[0, t'_0 + x] + \mu(\tau' + t) - \mu(\tau') \\ &\quad = \int_{\tau'}^{\tau'+t} \mathbb{I}_{\{t'_0+x\geq s\}} \nu_s[0, t'_0 + x - s] ds - \int_{\tau'}^{\tau'+t} m(u)\, du - \mu^0(\tau' + t) + \mu^0(\tau'), \end{aligned}$$

and for $s \geq \tau'$,

$$\mathbb{I}_{\{t'_0+x\geq s\}} \nu_s[0, t'_0 + x - s] \leq \mathbb{I}_{\{t'_0+x\geq s\}} \nu_s[0, t'_0 + x - \tau'] \leq \mathbb{I}_{\{t'_0+x\geq s\}} \nu_s[0, 2\delta_0] < c_0,$$

where the last inequality follows because $t'_0 + x < \tau' + 2\delta_0 < \tau' + 2$. The last two assertions, together with the definition of c_0 and the fact that μ^0 is non-decreasing, show that for any $x \in [0, \delta_0]$, $t \mapsto \alpha_t[0, t'_0 + x] - \mu(t)$ is non-increasing on $[\tau', \tau' + 2]$. □

We now state the main result of this section, whose proof is given in Section 4.1.4.

Theorem 4.10 *Suppose* (α, μ) *satisfies Assumption 4.5. Then the minimal solution* $(\xi, \beta, \iota, \rho)$ *of* (4.1) *is the unique solution of* (4.2) *in* $\mathbb{D}_{\mathcal{M}} \times \mathbb{D}_{\mathcal{M}}^{\uparrow} \times \mathbb{D}_{\mathbb{R}}^{\uparrow} \times \mathbb{D}_{\mathbb{R}}^{\uparrow}$.

In the next section we prove Theorem 4.10.

4.1.4 Proof of Theorem 4.10

Fix (α, μ) satisfying Assumption 4.5. In light of the uniqueness of a minimal solution established in Proposition 4.2, it suffices to show that a solution to (4.1) is minimal if and only if it satisfies condition (4.2)(iii). This is established in Propositions 4.11 and 4.12 below.

Proposition 4.11 *Suppose* (α, μ) *satisfies Assumption 4.5, and let* $(\xi, \beta, \iota, \rho)$ *be a solution of* (4.1)*. If* (ξ, ρ) *satisfy condition* (4.2)*(iii) then* $(\xi, \beta, \iota, \rho)$ *is a minimal solution of* (4.1)*.*

Proof: Let $(\xi, \beta, \iota, \rho)$ be a solution of (4.1). We will assume that ρ is not minimal and show that then

$$\gamma^{\rho}\left(\{t : \sigma(t) > t\}\right) > 0, \tag{4.5}$$

where recall that γ^{ρ} is the Lebesgue-Stieltjes measure associated with ρ, as defined in (1.3). This would then contradict (4.2)(iii), and hence prove the proposition. To this end, denote by $(\xi^*, \beta^*, \iota^*, \rho^*)$ the minimal solution of (4.1) and define

$$\varDelta(t) := \rho(t) - \rho^*(t), \qquad \text{and} \qquad \tau := \inf\{t \geq 0 : \varDelta(t) > 0\},$$

where we follow the convention that $\rho(0-) = \rho^*(0-) = \varDelta(0-) = 0$. Then the assumption that ρ is not minimal implies $\tau < \infty$. Also, provided $\tau > 0$, we have $\Delta(\tau-) = 0$ and the solutions $(\xi, \beta, \iota, \rho)$ and $(\xi^*, \beta^*, \iota^*, \rho^*)$ of (4.1) agree on $[0, \tau)$. Moreover, since (4.1)(i) implies $(\xi, \beta, \iota) = \Theta(\alpha, \mu + \rho)$ and $(\xi^*, \beta^*, \iota^*) = \Theta(\alpha, \mu + \rho^*)$, it follows from Lemma 2.7 that for any $x \geq 0$,

$$\xi[0, x] = \varGamma_1[\psi_x] \qquad \text{and} \qquad \xi^*[0, x] = \varGamma_1[\psi_x^*], \tag{4.6}$$

where for conciseness, we set

$$\psi_x(t) := \alpha_t[0, x] - \mu(t) - \rho(t), \qquad \text{and} \qquad \psi_x^*(t) := \alpha_t[0, x] - \mu(t) - \rho^*(t), \qquad t \geq 0. \tag{4.7}$$

We distinguish two mutually exhaustive cases.

Case 1: $\varDelta(\tau) = \varDelta(\tau-)$.
In this case $\rho(\tau) = \rho^*(\tau)$ and so the solutions agree on $[0, \tau]$. In particular, we have

$$\xi_\tau[0, x] = \xi_\tau^*[0, x], \quad x \geq 0. \tag{4.8}$$

Given that Assumption 4.5 holds, let $\delta_0 \in (0, 1)$ be as in Lemma 4.9 (with $\tau' = \tau$). Then we have the following claim.

Claim. If there exists $t_0 \in [\tau, \tau + \delta_0]$ and $x \in (0, \delta_0)$ such that $\xi_{t_0}[0, t_0 + x] = 0$ and $\gamma^{\rho}[t_0, t_0 + \varepsilon] > 0$ for some $\varepsilon > 0$, then (4.5) holds.

Proof of Claim. By the choice of δ_0, Lemma 4.9 (with $\tau' = \tau$, $t_0' = t_0$), (4.7) and the fact that ρ is non-decreasing imply that for every $x \in (0, \delta_0)$, $t \mapsto \psi_{t_0 + x}(t)$ is non-increasing on $[\tau, \tau + 2]$. For any such x, since $\xi_{t_0}[0, t_0 + x] = 0$, (4.6) and Lemma 2.3(2) together imply that

$\xi_t[0, t_0+x] = 0$ for all $t \in [t_0, \tau+2]$. But this implies that $\sigma(t) > t$ for every $t \in [t_0, t_0+x)$, and hence, (4.5) follows from the assumption of the claim that $\gamma^\rho[t_0, t_0+\varepsilon] > 0$ for some $\varepsilon > 0$. □

To complete the proof of (4.5) under Case 1, it suffices to verify the assumptions of the claim. To this end, let $t_2 \in (\tau, \tau+\delta_0/2)$ be such that $\Delta(t_2) > 0$ (such a t_2 exists by the definition and finiteness of τ), and let $t_1 := \inf\{t \in [\tau, t_2] : \rho(t) = \rho(t_2)\}$. Then, since $\Delta(\tau) = 0$, clearly t_1 is a strict maximizer of ρ on $[\tau, t_1]$, namely,

$$t \in [\tau, t_1) \text{ implies } \rho(t) < \rho(t_1). \tag{4.9}$$

By the right-continuity of ρ, the minimality of the solution $(\xi^*, \beta^*, \iota^*, \rho^*)$ and the fact that $\Delta(t_2) > 0$ and ρ^* is non-decreasing, we have $\rho(t_1) = \rho(t_2) > \rho^*(t_2) \geq \rho^*(t_1)$, and so $t_1 > \tau$. Denote $\kappa := \Delta(t_1) > 0$. For every $t \geq 0$, $\alpha_t \in \mathcal{M}_0$ by Assumption 4.5(i) and hence, it follows from the relation $(\xi, \beta, \iota) = \Theta(\alpha, \mu+\rho)$ and Proposition 2.8 that $\xi_t \in \mathcal{M}_0$. Together with the fact that α is right-continuous, we can find $\varepsilon \in (0, \delta_0/2)$ such that, with $y = t_1 - \varepsilon$ and $z = t_1 + \varepsilon$, we have $y \in (\tau, t_1)$ and

$$\xi_\tau(y, z] + \alpha^\tau_{t_1-\tau}(y, z] \leq \kappa/2, \tag{4.10}$$

where above and in what follows, we use the notation $f^T(\cdot) = f(T+\cdot) - f(T)$, $T > 0$, from (2.2). Fix such an $\varepsilon > 0$ and the corresponding y and z. We now compare $\xi_t[0, z]$ and $\xi^*_t[0, y]$ using the relations in (4.6) and (4.7). First note that

$$\xi_\tau[0, z] + \psi^\tau_z(t) = \xi_\tau[0, y] + \psi^{*,\tau}_y(t) + \xi_\tau(y, z] + \alpha^\tau_t(y, z] - \Delta(\tau+t), \quad t \geq 0,$$

where we used the fact that $\Delta(\tau) = \Delta(\tau-) = 0$. Substituting $t = t_1 - \tau$ and $\Delta(t_1) = \kappa$ above and using (4.10) and the fact that $\xi_\tau = \xi^*_\tau$, we obtain

$$\xi_\tau[0, z] + \psi^\tau_z(t_1-\tau) \leq \xi_\tau[0, y] + \psi^{*,\tau}_y(t_1-\tau) + \kappa/2 - \kappa = \xi^*_\tau[0, y] + \psi^{*,\tau}_y(t_1-\tau) - \kappa/2.$$

However, since the minimal solution satisfies (4.1)(ii) and $y < t_1$, we have $\xi^*_{t_1}[0, y] = 0$. When combined with (4.6) and Lemma 2.3(2), it follows that $\psi^{*,\tau}_y(t_1-\tau) \leq -\xi_\tau[0, y]$. Together with the last display, this means that

$$\xi_\tau[0, z] + \psi^\tau_z(t_1-\tau) \leq -\kappa/2. \tag{4.11}$$

Next, define

$$t_0 := \inf\{t \geq \tau : \xi_\tau[0, z] + \psi^\tau_z(t-\tau) \leq 0\}. \tag{4.12}$$

Then (4.11) and the fact that $t_1 \leq t_2 < \tau + \delta_0/2$ imply $t_0 \in [\tau, t_1] \subset [\tau, \tau+\delta_0]$ and from (4.12), it is clear that $\inf_{s \in [0, t_0-\tau]} \psi^\tau_z(s) = \psi^\tau_z(t_0-\tau) \leq -\xi_\tau[0, z]$. Thus, Lemma 2.3(2) implies that $\xi_{t_0}[0, z] = 0$. Now, $x := z - t_0$ lies in $[0, \delta_0]$ because $z = t_1 + \varepsilon$, $t_0 < t_1 \leq t_0 + \delta_0/2$ and $\varepsilon < \delta_0/2$. Thus, we have shown that $\xi_{t_0}[0, t_0+x] = 0$ for some $t_0 \in [\tau, \tau+\delta_0]$ and $x \in (0, \delta_0)$. To complete the verification of the assumptions of the claim, it suffices to show that γ^ρ charges $[t_0, t_1]$ (where the case $t_0 = t_1$ is possible), or equivalently, that $\rho(t_1) > \rho(t_0-)$. If $t_0 < t_1$ then this follows from (4.9). If $t_0 = t_1$ then by (4.11) and (4.12), ρ must have a jump at $t_0 = t_1$ (since $\psi_z - \rho = \alpha[0, z] - \mu$ is continuous by Assumption 4.5). Thus, $\rho(t_1) > \rho(t_1-)$ and so we have shown that γ^ρ charges the set $\{t \geq 0 : \sigma(t) > t\}$. This proves (4.5) for Case 1.

Case 2: $\Delta(\tau) > \Delta(\tau-)$.

In this case ρ must have a jump at τ (or, if $\tau = 0$, one must have $\rho(0) > 0$). Hence, it suffices to show that $\sigma(\tau) > \tau$. Consider first the case $\tau > 0$. In this case, let $c := \Delta(\tau) - \Delta(\tau-) = \rho(\tau) - \rho^*(\tau)$, and note that $c > 0$ by the case assumption. By (4.1)(ii), for every $y \in [0, \tau)$, $\xi^*_\tau[0, y] = 0$. The equation (4.7), with $x = y$, and Lemma 2.3(2) then imply that $\inf_{t\in[0,\tau]} \psi^*_y(t) = \psi^*_y(\tau) \leq -\xi^*_0[0, y]$. Since α_τ has no atoms, one can find y and z with $y < \tau < z$ such that $\alpha_\tau(y, z] < c$. Thus, recalling the definition of ψ_z in (4.7), we have

$$\psi_z(t) = \psi^*_y(t) + \alpha_t(y, z] - c\mathbb{I}_{\{t=\tau\}}, \quad t \in [0, \tau].$$

Since $\inf_{t\in[0,\tau]} \psi^*_y(t) = \psi^*_y(\tau)$ and $\alpha_t(y, z] - c\mathbb{I}_{\{t=\tau\}}$ is negative only when $t = \tau$, it follows that $\inf_{t\in[0,\tau]} \psi_y(t) = \psi_y(\tau) \leq \psi^*_y(\tau) \leq -\xi^*_0[0, y] = -\xi_0[0, y]$, where the last equality holds because $0 < \tau$. Another application of Lemma 2.3(2) in conjunction with (4.7) then shows that $\xi_\tau[0, z] = 0$. Since $z > \tau$, this implies $\sigma(\tau) > \tau$.

Finally, if $\tau = 0$, note that by (4.6), (4.7) and the explicit expression for Γ_1, for $z \geq 0$, $\xi^*_0[0, z] = \psi^{*,z}(0) \vee 0$, which is equal to $(\xi_{0-}[0, z] - \mu_0 - \rho^*(0)) \vee 0$, where ξ_{0-} is as in (4.3). Since $(\xi^*, \beta^*, \iota^*) = \Theta(\alpha, \mu + \rho^*)$ and $\alpha_0 = \xi_{0-}$ is absolutely continuous, by Proposition 2.8 ξ^*_0 has no atoms. Hence, $\xi^*_0[0, z] \to 0$ as $z \to 0$. Since $\rho(0) > \rho^*(0)$ (because $\tau = 0$) it follows that there exists $z > 0$ for which $\xi_0[0, z] = (\xi_{0-}[0, z] - \mu(0) - \rho(0)) \vee 0 = 0$. This shows that $\sigma(0) > 0$ and thus, proves (4.5) for Case 2. This completes the proof of the proposition. □

We now establish the converse result.

Proposition 4.12 *Suppose* (α, μ) *satisfies Assumption 4.5, and let* $(\xi, \beta, \iota, \rho)$ *be a solution of* (4.1) *for the data* (α, μ)*. If* $(\xi, \beta, \iota, \rho)$ *is a minimal solution of* (4.1)*, then* (ξ, ρ) *satisfies condition* (4.2)*(iii).*

Proof: We again proceed by proving the contrapositive. Fix (α, μ) that satisfies Assumption 4.5, and let $(\xi, \beta, \iota, \rho)$ be a solution of (4.1) for which (4.2)(iii) is false. The proof is established by showing that $(\xi, \beta, \iota, \rho)$ is not minimal by explicitly constructing another solution $(\tilde{\xi}, \tilde{\beta}, \tilde{\iota}, \tilde{\rho})$ of (4.1) for which $\rho \leq \tilde{\rho}$ is false. First, note that (4.1)(i) and Lemma 2.7 imply that

$$\xi[0, x] = \Gamma_1(\psi_x), \quad \text{where} \quad \psi_x := \alpha[0, x] - \mu - \rho, \quad x \geq 0. \tag{4.13}$$

We will find it convenient to use the following equivalent form of (4.2)(iii):

$$\{\sigma_t = t \;\, d\rho\text{-a.e.}\} \Longleftrightarrow \{\forall\, \delta > 0,\; \xi_t[0, t + \delta] > 0 \;\, d\rho\text{-a.e.}\}.$$

Since, by our assumptions, (4.2)(iii) does not hold, there exist $\delta > 0$ and a measurable set $B \subset \{t \geq 0 : \sigma(t) \geq t + \delta\}$ with $\gamma^\rho(B) > 0$. Assume without loss of generality that B is bounded, and denote by T the essential supremum of the restriction of γ^ρ to B:

$$T := \sup\{t \in [0, \infty) : \gamma^\rho(B \cap [t, \infty)) > 0\}.$$

Then $T \in [0, \infty)$ and we must have $\gamma^\rho(B \cap [0, T]) > 0$. We now distinguish two mutually exclusive and exhaustive cases.

Case 1. $T \notin B$ or $\gamma^\rho(\{T\}) = 0$.
Since $\gamma^\rho(B \cap [0, T]) > 0$, the assumptions of this case then imply $T > 0$ and for every $t \in [0, T)$, there exists $t_0 \in [t, T)$ such that

$$\sigma(t_0) \geq t_0 + \delta \quad \text{and} \quad \rho(T-) > \rho(t_0). \tag{4.14}$$

Fix $t_0 \in (T-\delta, T)$ for which (4.14) holds and choose $y \in (T, t_0+\delta)$. Then we have

$$0 < T - \delta < t_0 < T < y < t_0 + \delta. \tag{4.15}$$

Also, because $\sigma(t_0) \geq t_0 + \delta$ and $y < t_0 + \delta$, the fact that ξ satisfies (4.2)(ii) implies

$$\xi_{t_0}[0, y] = 0. \tag{4.16}$$

Moreover, let δ_0 be the quantity in Lemma 4.9 when $\tau' = t_0$ and without loss of generality assume that $\delta < \delta_0$. Then we can set $t_0' = T$ and $x = y - T$ in Lemma 4.9 to conclude that

$$t \mapsto \alpha_t[0, y] - \mu(t) \text{ is non-increasing on } [t_0, t_0 + 2]. \tag{4.17}$$

We now construct $\tilde{\rho} \in \mathbb{D}_{\mathbb{R}}^{\uparrow}$ as follows:

$$\tilde{\rho}(t) := \begin{cases} \rho(t), & t \in [0, t_0), \\ \rho(t_0), & t \in [t_0, T), \\ \rho(t), & t \in [T, \infty). \end{cases}$$

Let $(\tilde{\xi}, \tilde{\beta}, \tilde{\iota}) = \Theta(\alpha, \mu + \tilde{\rho})$, and note that then, again by Lemma 2.7, we have the analog of (4.13):

$$\tilde{\xi}[0, x] = \Gamma_1(\tilde{\psi}_x), \qquad \text{where} \qquad \tilde{\psi}_x := \alpha[0, x] - \mu - \tilde{\rho}, \qquad x \geq 0. \tag{4.18}$$

Our goal now is to show that (4.1)(ii) holds for $\tilde{\xi}$; once this is established, one has a solution $(\tilde{\xi}, \tilde{\beta}, \tilde{\iota}, \tilde{\rho})$ of (4.1) with $\tilde{\rho}(T-) = \rho(t_0) < \rho(T-)$, where the last inequality is due to (4.14), thus contradicting the minimality of the solution $(\xi, \beta, \iota, \rho)$ of (4.1).

To show that (4.1)(ii) holds for $\tilde{\xi}$ or, equivalently, that $\tilde{\xi}_t[0, t) = 0$ for all $t > 0$, first note that when $t \in [0, t_0]$, this follows from the corresponding property for ξ because ρ and $\tilde{\rho}$, and hence, by (4.13) and (4.18), ξ and $\tilde{\xi}$, coincide on $[0, t_0]$. Next, consider $t \geq T$ and fix $z < t$. Showing (4.1)(ii) for $\tilde{\xi}$ here amounts to showing that for any $z < t$, $\tilde{\xi}_t[0, z] = 0$. Since ξ satisfies (4.1)ii), we know that $\xi_t[0, z] = 0$ for such t and z. Together with (4.13) and Lemma 2.3(2), this implies that $\psi_z(t) = \inf_{s \in [0,t]} \psi_z(s) \leq -\xi_0[0, z]$. When combined with the relations $\rho(t) = \tilde{\rho}(t)$, $\rho(s) \geq \tilde{\rho}(s)$ for all $s \in [0, t]$ and $\xi_0 = \tilde{\xi}_0$, we see that $\inf_{s \in [0,t]} \tilde{\psi}_z(s) = \tilde{\psi}_z(t) = \psi_z(t) \leq -\tilde{\xi}_0[0, z]$. Due to (4.18) and Lemma 2.3(2), the last relation shows that $\tilde{\xi}_t[0, z] = 0$.

Finally, we consider $t \in (t_0, T)$ and establish a stronger claim, namely, that $\tilde{\xi}_t[0, y] = 0$ (recall that $y > T$). In this case, since $\xi_{t_0} = \tilde{\xi}_{t_0}$, (4.16) implies that $\tilde{\xi}_{t_0}[0, y] = 0$. Moreover, since $\tilde{\rho}$ is non-decreasing, (4.17) implies that $\alpha[0, y] - \mu - \tilde{\rho}$ is non-increasing on $[t_0, t_0 + 2]$. Together with (4.18) and Lemma 2.3(2) this implies that $\tilde{\xi}_t[0, y] = 0$ for $t \in [t_0, t_0 + 2]$ and in particular, for all $t \in [t_0, T]$. As a result, $\tilde{\xi}_t[0, t) = 0$ for all $t \geq 0$, which implies $\tilde{\xi}$ satisfies (4.1)(ii) as claimed.

Case 2: $T \in B$ and $\gamma^\rho(\{T\}) > 0$.

In this case, $\sigma(T) \geq T + \delta$ by the definition of B. Setting $\rho(0-) = 0$, for an arbitrary $T_1 > T$, we define

$$\tilde{\rho}(t) := \begin{cases} \rho(t), & t \in [0, T), \text{ if } T > 0, \\ \rho(T-), & t \in [T, T_1), \\ \rho(t), & t \in [T_1, \infty). \end{cases}$$

Since $\rho \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$, clearly $\tilde{\rho}$ also lies in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$. Define $(\tilde{\xi}, \tilde{\beta}, \tilde{\iota}) := \Theta(\alpha, \mu + \tilde{\rho})$ and, as in Case 1, note that (4.18) holds. By construction, $\tilde{\rho}(t) \le \rho(t)$ for every $t \in [0, \infty)$ and for $t \in [T, T_1)$, $\tilde{\rho}(t) = \rho(T-) < \rho(T) \le \rho(t)$, where we used the case assumption, $\gamma^{\rho}(\{T\}) > 0$. Therefore, the proof will be complete if we can show that $\tilde{\xi}$ satisfies (4.1)(ii), that is, $\tilde{\xi}_t[0, t) = 0$ for all $t \ge 0$. The proofs of this equality for the cases $t \in [0, T)$ and $t \in [T_1, \infty)$ follow exactly as in Case 1.

For the intermediate case, fix $t \in [T, T_1)$ and $y < t$. It remains to show that $\tilde{\xi}_t[0, y] = 0$. Observe that since $\xi_t[0, t) = 0$ for all $t \ge 0$ by (4.1)(ii) and $y < t < T_1$, we have in particular that $\xi_t[0, y] = \xi_{T_1}[0, y] = 0$. Therefore, Lemma 2.3(2) and (4.13) imply

$$\inf_{s \in [0, T_1 - t]} \psi^t_y(s) = \psi^t_y(T_1 - t) \le 0, \tag{4.19}$$

where recall the notation $f^T(\cdot) = f(T + \cdot) - f(T)$ from (2.2). We now show that the relation (4.19) also holds when ψ_y is replaced everywhere by $\tilde{\psi}_y$. This would conclude the proof of Case 2 because then, due to the already verified property that $\tilde{\xi}_{T_1}[0, y] = 0$ and (4.18), another application of Lemma 2.3(2) would imply that $\tilde{\xi}_t[0, y] = 0$. To this end, we write

$$\tilde{\psi}^t_y(s) = \psi^t_y(s) + \rho^t(s) - \tilde{\rho}^t(s), \qquad s \in [0, T_1 - t]. \tag{4.20}$$

By definition, $\tilde{\rho}(T_1) = \rho(T_1)$, and so $\rho^t(T_1 - t) - \tilde{\rho}^t(T_1 - t) = \tilde{\rho}(t) - \rho(t) \le 0$. Thus, $\tilde{\psi}^t_y(T_1 - t) \le \psi^t_y(T_1 - t)$ which, together with (4.19), implies

$$\inf_{s \in [0, T_1 - t]} \tilde{\psi}^t_y(s) \le \tilde{\psi}^t_y(T_1 - t) \le \psi^t_y(T_1 - t) = \inf_{s \in [0, T_1 - t]} \psi^t_y(s) \le 0. \tag{4.21}$$

To conclude the proof, we show that the first inequality in (4.21) can be replaced by equality. Lemma 2.3(1), the explicit expression for Γ_1 in (2.1) and the first inequality in (4.21) imply $\tilde{\xi}_{T_1}[0, y] = \Gamma_1(\tilde{\xi}_t[0, y] + \tilde{\psi}^t)(T_1 - t) = \tilde{\psi}^t_y(T_1 - t) - \inf_{s \in [0, T_1 - t]} \tilde{\psi}^t_y(T_1 - s)$. Since we showed above that $\tilde{\xi}_{T_1}[0, y] = 0$, this completes the proof of Case 2, and hence of the proposition. □

4.1.5 An Optimality Result

The fact established in Theorem 4.10, that our fluid model for the hard EDF, (4.2), minimizes ρ among all solutions of (4.1), may be interpreted as follows: *Assume hard EDF is employed and jobs may renege prior to or at their deadlines. Then reneging only at the time of the deadline minimizes the reneging count at all times.* We now argue that another optimality property may be deduced from Theorem 4.10, one that has the following interpretation. *Assume an arbitrary service policy is employed, and reneging prior to or at the deadline is allowed. Then employing the hard EDF policy and reneging exactly when the deadline elapses minimizes the reneging count at all times.*

A precise statement is as follows. Given $(\alpha, \mu) \in \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$, a tuple $(\xi, \beta, \iota, \rho) \in \mathbb{D}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathcal{M}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}} \times \mathbb{D}^{\uparrow}_{\mathbb{R}}$ is said to be *compatible with the data* (α, μ) if it adheres to items 1 and 4 of Definition 2.5. with $\mu + \rho$ substituted for μ, and to (4.1)(ii). That is,

$$\begin{cases} \xi[0, x] = \alpha[0, x] - \mu - \rho + \beta(x, \infty) + \iota, \text{ for every } x, \\ \beta[0, \infty) + \iota = \mu + \rho, \\ \xi_t[0, t) = 0, \text{ for every } t > 0. \end{cases} \tag{4.22}$$

This gives a model for an arbitrary service policy, that does not necessarily respect the priority determined by deadlines, as required in assumption 2 of Definition 2.5, and need not be non-idling as required by assumption 3 of that definition. Finally, the reneging need not adhere to the rule 'renege exactly at the time of the deadline.' as required by (*iii*) of Definition 4.2.

Proposition 4.13 *Let* (α,μ) *satisfy Assumption 4.5. Let* (ξ,β,ι,ρ) *be compatible with the data. Let also* $(\xi^0,\beta^0,\iota^0,\rho^0)$ *denote the unique solution of* (4.2)*. Then* $\rho^0 \le \rho$.

Proof: Since (α,μ) satisfy Assumption 4.5, Theorem 4.10 is applicable. Thus, it suffices to argue that there exists a triplet (ξ^1,β^1,ι^1) such that $(\xi^1,\beta^1,\iota^1,\rho)$ forms a solution of (4.1), since by Theorem 4.10 this would give $\rho^0 \le \rho$. To construct such a triplet, simply set $(\xi^1,\beta^1,\iota^1) = \Theta(\alpha,\mu+\rho)$. It remains to show (4.1)(ii), namely $\xi^1_t[0,t)=0$ for every $t>0$. To this end we appeal to the minimality property of the SM on the half-line (see Section 2 of [6]), which states the following: *If* $(\varphi,\eta)\in\mathbb{D}_{\mathbb{R}}\times\mathbb{D}^{\uparrow}_{\mathbb{R}}$ *and* $\varphi(t)+\eta(t)\ge 0$ *for all* t*, then*

$$\Gamma_1[\varphi](t) \le \varphi(t)+\eta(t), \quad t\ge 0.$$

Given x, let $\varphi=\alpha[0,x]-\mu-\rho$ and $\eta=\beta(x,\infty)+\iota$. Then $\xi_t[0,x]=\varphi(t)+\eta(t)\ge 0$ for all t, by (4.22). Next, by (2.8), $\xi^1_t[0,x]=\Gamma_1[\alpha[0,x]-\mu-\rho](t)$. As a result, $\xi^1_t[0,x]\le\xi_t[0,x]$. Since x and t are arbitrary, and we have the identity $\xi_t[0,t)=0$ by (4.22), it follows that $\xi^1_t[0,t)=0$ for all t. This completes the proof. □

4.2 Fluid Models for Policies that Prioritize by Job Size

We now describe two variants of a scheduling policy where priority is determined by the job size or processing requirement, where by 'processing requirement' one refers to the time it takes a server, when operating at unit rate, to complete processing of the job. In both of these systems, jobs arrive into an infinite buffer served by a single server, with their processing requirements known in advance. The server works according to a rule that, at any time, gives priority to the job that has the smallest processing requirement. As mentioned in the introduction, the non-preemptive version of the policy, where the service of a job is not interrupted by the arrival of a new job (that has a smaller size), is referred to as *shortest job first* (SJF) and the preemptive version of the policy is called *shortest remaining processing time* (SRPT).

The description of the data for the fluid model is quite similar to that of the FIFO discipline discussed in Section 3.3, except that we now take the mass to have the meaning of amount of work, rather than the number of jobs arrived. More precisely, as in Section 3.2, we suppose that we are given a measurable locally integrable function $\lambda:\mathbb{R}_+\times\mathbb{R}_+\to\mathbb{R}_+$ that admits the following interpretation: during the time interval $(t,t+dt)$, $\lambda(t,y)dydt$ jobs arrive with size in the interval $(y,y+dy)$. Expressed in terms of work, we can say that $y\lambda(t,y)dydt$ represents the amount of work that arrived in the interval $(t,t+dt)$, due to jobs with size in $(y,y+dy)$. Thus, the total arrived workload of jobs of different sizes is captured by the measure-valued path α, defined by

$$\hat{\alpha}_t[0,x]=\int_{[0,t]\times[0,x]} y\lambda(s,y)dsdy,\ t\in\mathbb{R}_+, x\in\mathbb{R}_+.$$

As before, we assume that the distribution of mass in the queue in terms of job sizes prior to zero is captured by the measure ξ_{0-} and let $\alpha=\xi_{0-}+\hat{\alpha}$, and we also assume that we are given $\mu \in \mathbb{D}_{\mathbb{R}}^{\uparrow}$, where $\mu(t)$ denotes the potential amount of work that the server can process in the interval $[0, t]$. Denote by $\xi_t[0, x]$ the amount of work in the buffer, due to jobs whose processing requirements lie within $[0, x]$, and let $\beta_t[0, x]$ represent the amount of work (and not number of jobs) processed by the server for the same class of jobs. Then, we expect the fluid models for both SJF and SRPT to satisfy equation (3.7). The equations that describe the probabilistic model are presented in Section 5.2. As shown there, the state descriptors for the stochastic SJF model satisfy the same relation in terms of Θ; see (5.44). This makes the state descriptor for the workload in the SJF model particularly easy to analyze, although, as shown in Section 4.2, the proof of convergece of the state of the number of jobs in the SJF system is considerably more involved. In the case of SRPT, additional considerations are required to deal with a certain error term.

5 Convergence and characterization of limits

We now use the tools introduced above to describe the queueing models associated with three scheduling policies, and establish convergence of the queueing model under the LLN scaling to the fluid models described in Section 4. The EDF policy is considered in Section 5.1 and the SJF and SRPT policies in Section 5.2, respectively.

5.1 Earliest-Deadline-First Convergence Results

In Section 5.1.1, we introduce the primitive processes that describe the stochastic hard EDF model, and form the equations governing the dynamics. The latter are analogous, but not identical, to the fluid model equations introduced in Section 4.1.1. In Section 5.1.2 we introduce the fluid scaling and state the main convergence result, Theorem 5.4. The proof of Theorem 5.4, which is given in Section 5.1.4, builds on tightness results that are established in Section 5.1.3. The soft EDF model is easier to analyze using our MVSP. Indeed, as explained in Remark 5.6, convergence of the sequence of scaled stochastic soft EDF models to its corresponding fluid limit also follows as an immediate corollary of Theorem 5.4.

5.1.1 Equations Governing the Stochastic Model

We recall the verbal description of the EDF queueing model given in Section 4.1.1. To describe its dynamics precisely, let the scaling parameter be denoted by $N \in \mathbb{N}$; we refer to the queueing model corresponding to N as the *N-system*, or, for simplicity, the *system*. The random variables and stochastic processes introduced below are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The model primitives that determine the dynamics of the N-system consist of a measure-valued arrival process $\hat{\alpha}^N$, real-valued processes S and μ^N, that together describe the service, and a measure ξ_{0-}^N that captures the state of the buffer just prior to zero. For $t, x \geq 0$, let $\hat{\alpha}_t^N[0, x]$ denote the number of jobs that have arrived during the time interval $[0, t]$ with deadlines in $[0, x]$. This does not include jobs that are counted in the measure ξ_{0-}^N, where $\xi_{0-}^N[0, x]$ represents the jobs present in the buffer at time 0 (not counting the job in service) with deadlines in $[0, x]$. We shall assume that for each t, $\hat{\alpha}_t^N[0, t) = 0$, meaning that jobs cannot have (absolute) deadlines that are smaller than their time of arrival. We let

$$\alpha^N = \hat{\alpha}^N + \xi_{0-}^N. \tag{5.1}$$

The model for service is based on two stochastic elements: the integer-valued potential service process S (independent of N) that captures the service requirements of jobs, and the cumulative effort process μ^N that allows for variable rate of service, both of which have sample paths in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$. Specifically, the process S is assumed to be a non-delayed renewal counting process with inter-renewal times distributed according to the service times of jobs. We assume that the inter-renewal distribution of S has mean 1 (there is no loss of generality because of the way we will employ the process μ^N, as explained below). By assumption, $S(0) = 1$, and given $t \geq 0$, $S(t) - 1$ represents the number of jobs completed by the time the server has been occupied for t units of time, assuming service is provided at rate 1. Let B^N be a càdlàg $\{0, 1\}$-valued process describing the state of the server, namely,

$$B^N(t) := \begin{cases} 1 & \text{if the server is busy at time } t, \\ 0 & \text{otherwise,} \end{cases}$$

and let $B^N(0-)$ be the initial state of the server. We allow the rate of service to vary over time, and so the actual number of job completions by time t is given by $S(T^N(t)) - 1$, where

$$T^N(t) := \int_{[0,t]} B^N(s)d\mu^N(s), \qquad t \geq 0, \tag{5.2}$$

represents the cumulative effort spent by the server in $[0, t]$.

The state of the buffer is described by the process ξ^N, which has sample paths in $\mathbb{D}_{\mathcal{M}}$. Analogous to ξ^N_{0-}, for $t, x \geq 0$, $\xi^N_t[0, x]$ represents the number of jobs that are in the buffer at time t (not counting the job in service) and have deadline within $[0, x]$. Note that the total number of jobs in the system at time t (including those in the queue and the one in service) is then given by $\xi^N_t[0, \infty) + B^N(t)$. The left end of the support of ξ^N_t will play an important role in the analysis. We denote

$$\sigma^N(t) := \min \operatorname{supp}[\xi^N_t], \qquad t \geq 0. \tag{5.3}$$

Auxiliary processes that help describe the dynamics of the system are the measure-valued processes $\beta^{s,N}$, $\beta^{r,N}$, β^N, all of whom have sample paths in $\mathbb{D}^{\uparrow}_{\mathcal{M}}$, and the real-valued processes ρ^N and ι^N. For $t, x \geq 0$, the cumulative number of jobs with deadline in $[0, x]$ that started service (and possibly departed from the system) before time t is given by $\beta^{s,N}_t[0, x]$, and those with deadline in $[0, x]$ that reneged from the system before time t because their deadlines elapsed before they could be admitted into service is given by $\beta^{r,N}_t[0, x]$. If we set

$$\beta^N = \beta^{s,N} + \beta^{r,N}, \tag{5.4}$$

then $\beta^N_t[0, x]$ represents the total number of jobs with deadlines in $[0, x]$ that have left the buffer by time t. The reneging count process is denoted by ρ^N and has sample paths in $\mathbb{D}^{\uparrow}_{\mathbb{R}}$. For $t \geq 0$, $\rho^N(t)$ is the total number of jobs that have reneged in the time interval $[0, t]$, namely

$$\rho^N(t) = \beta^{r,N}_t[0, \infty) = \beta^{r,N}_t[0, t], \qquad t \geq 0, \tag{5.5}$$

where the last equality captures the fact that jobs in the buffer (that are still awaiting service) renege only when the current time exceeds their deadline. In particular, this implies

$$\beta^{r,N}_t[0, x] = \rho^N(t \wedge x), \qquad t, x \geq 0, \tag{5.6}$$

and thus the measure-valued process $\beta^{r,N}$ can be recovered from the real-valued process ρ^N. Moreover, the total number of jobs sent to service by time t satisfies

$$\beta_t^{s,N}[0,\infty) = S(T^N(t)) - 1 + B^N(t), \qquad t \geq 0. \tag{5.7}$$

Next, analogous to the process T^N defined in (5.2), we let

$$\iota^N(t) := \int_{[0,t]} (1 - B^N(s)) d\mu^N(s) = \mu^N(t) - T^N(t), \quad t \geq 0. \tag{5.8}$$

In the special case $\mu_t^N = t$, $t \geq 0$, the process ι^N represents the cumulative idle time of the server; in general it is the total lost service effort due to idleness. Finally, it will be useful to denote

$$e^N(t) := \beta_t^{s,N}[0,\infty) - T^N(t), \quad t \geq 0. \tag{5.9}$$

In view of (5.2), (5.7) and the fact that S has mean 1, it is apparent that $e^N(t)/N$ will play the role of an error term.

We now write several identities that follow directly from the above description of the processes and the EDF policy. In these equations, $x, t \in \mathbb{R}_+$ are arbitrary. First, note that

$$\xi_t^N[0,x] = \alpha_t^N[0,x] - \beta_t^{s,N}[0,x] - \rho^N(t \wedge x), \tag{5.10}$$

$$\beta_t^{s,N}[0,x] = \mu^N(t) - \beta_t^{s,N}(x,\infty) - \iota^N(t) + e^N(t), \tag{5.11}$$

where the first is the balance equation for jobs with deadline in $[0,x]$, and the second is immediate from (5.8) and (5.9). Now, (5.5) and (5.6) imply that $\beta_t^{r,N}(x,\infty) = \rho^N(t) - \rho^N(t \wedge x)$. Combining this with (5.10), (5.11) and (5.4), we obtain

$$\xi_t^N[0,x] = \alpha_t^N[0,x] - \mu^N(t) - \rho^N(t) - e^N(t) + \beta_t^N(x,\infty) + \iota^N(t). \tag{5.12}$$

Sending $x \to \infty$ in (5.12), we also have

$$\xi_t^N[0,\infty) = \alpha_t^N[0,\infty) - \mu^N(t) - \rho^N(t) - e^N(t) + \iota^N(t). \tag{5.13}$$

Next, the EDF priority rule dictates that when a job is sent to the server, no job in the queue has a smaller deadline. Moreover, the non-idling property of the server implies that when the server idles no jobs are present in the buffer. These facts can be expressed by the relations

$$\int_{[0,\infty)} \xi_t^N[0,x] d\beta_t^{s,N}(x,\infty) = 0, \qquad x \geq 0, \tag{5.14}$$

$$\int_{[0,\infty)} \xi_t^N[0,\infty) d\iota^N(t) = 0, \tag{5.15}$$

where the integral in (5.14) is with respect to the t-variable, for a fixed x. By (5.4), (5.5), (5.8) and (5.9),

$$\beta_t^N[0,\infty) + \iota^N(t) = \mu^N(t) + \rho^N(t) + e^N(t). \tag{5.16}$$

Moreover, the reneging behavior of jobs is such that at any given time t, no jobs with deadline less than or equal to t are in the queue; and jobs that renege do so exactly at the time of their deadline. These two facts imply the identities

$$\xi_t^N[0,t] = 0, \tag{5.17}$$

$$\int_{[0,\infty)} \mathbb{I}_{\{\sigma^N(t-)>t\}} d\rho^N(t) = 0. \tag{5.18}$$

Note that we can deduce that (5.14) holds for $\beta^{r,N}$ as well. Indeed, fix x. It follows from (5.5) and (5.6) that $\beta_t^{r,N}(x,\infty) = \rho^N(t) - \rho^N(t \wedge x)$, and so the measure $d\beta_t^{r,N}(x,\infty)$ charges only a subset of the form $\{t_k\}$ of (x,∞). For each such t_k, $\xi_{t_k}^N[0,x] = 0$ by (5.17), since $t_k > x$. Thus (5.14) is valid for $\beta^{r,N}$. Since $\beta^N = \beta^{s,N} + \beta^{r,N}$ by (5.4), we have

$$\int_{[0,\infty)} \xi_t^N[0,x] d\beta_t^N(x,\infty) = 0. \tag{5.19}$$

Remark 5.1 An observation that will be useful in establishing the fluid limit theorem is that equations (5.10)–(5.19) are closely related to the fluid model equations (4.2). Indeed, comparing equations (5.12), (5.19), (5.15) and (5.16) with properties 1–4 in Definition 2.9 of the MVSP , and noting that $\mu^N + \rho^N + e^N$ is non-decreasng by (5.16), it follows that

$$(\xi^N, \beta^N, \iota^N) = \Theta(\alpha^N, \mu^N + \rho^N + e^N). \tag{5.20}$$

This is analogous to the fluid model equation (4.2)(i), except for the presence of the additional error term e^N. Further, (5.17) is the exact analog of equation (4.2)(ii), and (5.18) is similar to (4.2)(iii), with the notable difference of having $\sigma^N(t-)$ in the former and $\sigma(t)$ in the latter.

5.1.2 The EDF Fluid Limit Theorem

For measure-valued processes $\zeta = \alpha, \beta, \beta^s, \beta^r, \xi$ and real-valued processes $\gamma = \mu, \iota, \rho, e$, set

$$\bar{\zeta}_t^N(B) := \frac{\zeta_t^N(B)}{N}, \quad B \in \mathcal{B}(\mathbb{R}_+); \qquad \bar{\gamma}^N(t) := \frac{\gamma^N(t)}{N}, \quad t \geq 0. \tag{5.21}$$

There is no need to define a new version of the process σ^N defined in (5.3), because this process plays the same role for the scaled processes, in the sense that $\sigma^N(t) = \min \operatorname{supp}[\bar{\xi}_t^N]$, $t \geq 0$.

As observed in Remark 5.1, the stochastic model (and therefore its scaled version) satisfies equations that are close to the equations in (4.2). By Theorem 4.10, the latter characterize the minimal solution of the fluid model equations (4.1) when the fluid primitives α and μ satisfy Assumption 4.5. Thus, we now impose fairly general assumptions on the scaled stochastic primitives $\bar{\alpha}^N$ and $\bar{\mu}^N$ that ensure that their limits satisfy Assumption 4.5. Recall that the symbol '$\Rightarrow$' denotes convergence in distribution. Specifically, if π^N and π are $\mathbb{D}_{\mathcal{M}}$-valued random variables, $\pi^N \Rightarrow \pi$ means convergence in distribution in the Skorohod topology on càdlàg functions over $(\mathcal{M}, d_{\mathcal{L}})$. We now state our assumptions.

Assumption 5.2 *The following properties hold:*

1. *The sequence $\{\bar{\alpha}^N\}$ converges in distribution to α, where α is a (non-random) member of $\mathbb{C}^{\uparrow}_{\mathcal{M}_0}$ that satisfies Assumption 4.5(i).*
2. *The sequence $\{\bar{\mu}^N\}$ converges in distribution to μ, where μ is a (non-random) element of $\mathbb{C}^{\uparrow}_{\mathbb{R}}$ that has the form* (4.4).

Example 5.3 We provide some simple examples where the above assumption holds. For simplicity, the initial condition is set to zero in these examples, so $\bar{\alpha}^N = N^{-1}\hat{\alpha}^N$.
(a) First, consider a time homogeneous setting. In the Nth system, arrivals follow a renewal

process $S^N_{\rm arr}$, that is an accelerated version of a fixed renewal process $S_{\rm arr}$, that is, $S^N_{\rm arr} = S_{\rm arr}(N\cdot)$. The interarrival distribution of $S_{\rm arr}$ has finite mean denoted $1/\lambda_0$. The patience of job i, denoted P_i, is assumed to be drawn from an iid sequence that follows a distribution ν^0. The potential service process, S, is a renewal process, modeling a fixed service time distribution. With $\{\tau_i\} = \{\tau^N_i\}$ denoting the jump times of $S^N_{\rm arr}$, $\hat\alpha^N$ is given by

$$\hat\alpha^N_t(dx) = \sum_{i=1}^{\infty} \mathbb{I}_{\{t\ge\tau_i\}}\delta_{P_i}(dx).$$

If ν^0 is atomless, then the convergence of $\bar\alpha^N$ to $\alpha \in \mathbb{C}^{\uparrow}_{\mathcal{M}_0}$ follows from the LLN, where $\hat\alpha$ takes the form

$$\hat\alpha_t[0,x] = \int_0^t \mathbb{I}_{\{x\ge s\}}\lambda_0\nu^0[0,x-s]ds.$$

Hence, (4.3) holds with $\nu_t = \lambda_0\nu^0$.

As for the service model, one can set $\mu_t = t$ for all t (in (4.4) this can be achieved by setting $\mu^0(t) = 0$, $m(t) = 1$ for all t), by which the assumption on μ clearly holds.

(b) A slight modification of (a) is to let $\{P_i\}$ still be an independent sequence, but not necessarily identically distributed. We assume here that the distributions of the P_i's alternate periodically within a finite collection of atomless distributions $\{\nu^k\}_{k=1}^K$. Then it is clear that the same conclusions hold with ν^0 replaced by $K^{-1}\sum_{k=1}^K \nu^k$.

(c) Next, we give an example where the Nth system's parameters vary periodically with period T, leading to a limit (α,μ) whose time derivative also varies periodically with period T. To this end, let $L \in \mathbb{N}$ and fix $0 = t_0 < t_1 < \cdots < t_L = T$ and, for $t \in \mathbb{R}_+$, denote by $f(t)$ the unique $t' \in [0,T)$ such that $nT + t' = t$, for some $n \in \mathbb{N}$. Assume that $S^N_{\rm arr}(t) = S_{\rm arr}(N\int_0^t \sum_l \theta_l \mathbb{I}_{\{f(s)\in[t_{l-1},t_l)\}}ds)$, where $S_{\rm arr}$ is as in example (a) and $(\theta_l)_{1\le l\le L}$ are positive constants. Thus the interarrivals within $[t_{l-1},t_l)$ have mean $1/(\lambda^0\theta_l)$. To allow also the patience distribution to be piecewise constant in a similar fashion, consider L iid sequences $\{P^l_i\}_{i\in\mathbb{N},1\le l\le L}$, where P^l_1 is distributed according to some atomless ν^l. We assume that jobs arriving within $[t_{l-1},t_l)$ have patience drawn from ν^l. This leads to the following model for $\hat\alpha^N$:

$$\hat\alpha^N_t(dx) = \sum_{i=1}^{\infty} \mathbb{I}_{\{t\ge\tau_i\}} \sum_{l=1}^{L} \mathbb{I}_{\{f(t)\in[t_{l-1},t_l)\}}\delta_{P^l_i}(dx).$$

Here, τ_i are again the jump times of $S^N_{\rm arr}$. It is a simple exercise to show that the corresponding $\bar\alpha^N$ converge to $\hat\alpha$ given by (4.3), where now $\nu_t = \lambda^0\sum_l \mathbb{I}_{\{f(t)\in[t_{l-1},t_l)\}}\nu^l$. We can similarly let the service time distribution vary over time by modifying μ. For example, we can take $\mu_t = \int_0^t \sum_l m_l \mathbb{I}_{\{f(s)\in[t_{l-1},t_l)\}}ds$, for some positive constants $(m_l)_{1\le l\le L}$.

Theorem 5.4 *Suppose Assumption 5.2 holds, and for the associated* (α,μ)*, let* (ξ,β,ι,ρ) *denote the unique solution of* (4.2) *(equivalently, the minimal solution of* (4.1)*). Then* (ξ,β,ι,ρ) *lies in* $\mathbb{C}_{\mathcal{M}_0}\times\mathbb{C}^{\uparrow}_{\mathcal{M}_0}\times\mathbb{C}^{\uparrow}_{\mathbb{R}}\times\mathbb{C}^{\uparrow}_{\mathbb{R}}$ *and* $(\bar\xi^N,\bar\beta^N,\bar\iota^N,\bar\rho^N)\Rightarrow(\xi,\beta,\iota,\rho)$.

Remark 5.5 Since the limits are continuous, the convergence stated above holds also in the u.o.c. topology.

Remark 5.6 Theorem 5.4 also implies convergence to the fluid limit under the soft EDF policy. To see why, consider a queueing model operating under the hard EDF policy over a time horizon $[0, T]$. If we add the constant T to all deadlines (of jobs initially in the system as well as those that arrive during the interval $[0, T]$) then there is no reneging (that is, $\rho \equiv 0$) and the hard and soft versions of the policy give rise to exactly the same state dynamics. Hence, we obtain convergence of the sequence of fluid scaled soft EDF models to the limit given by $(\xi, \beta, \iota) = \Theta(\alpha, \mu)$.

An outline of the proof is as follows. We begin in Section 5.1.3 by showing that the sequence of rescaled versions of $\Upsilon^N = (\alpha^N, \mu^N, \rho^N, e^N)$ is tight, and that the scaled error term e^N vanishes. Then, in Section 5.1.4, we show that given any convergent subsequence with limit $(\alpha, \mu, \rho, 0)$, the continuity of the MVSM established in Lemma 2.10 and the representation (5.20) together show that the rescaled versions of the corresponding $(\xi^N, \beta^N, \iota^N)$ converge to $\Theta(\alpha, \mu+\rho)$, thus establishing (4.2)(i). To show uniqueness of the limit, we then show that the remaining properties of (4.2) are also satisfied and invoke the uniqueness stated in Theorem 4.10. Relation (4.2)(ii) essentially follows on taking limits in (5.17). Limits in (5.18) do not automatically yield (4.2)(iii), and the proof of this requires additional estimates on the reneging process.

5.1.3 Tightness Results for the EDF Model

Recall that a sequence of processes with sample paths in $\mathbb{D}_{\mathcal{S}}$, $\mathcal{S}$ being a Polish space, is said to be *C-tight* if it is tight and, in addition, any subsequential limit has, with probability 1, paths in $\mathbb{C}_{\mathcal{S}}$.

To establish tightness, we will appeal to the following characterization of C-tightness of processes with sample paths in $\mathbb{D}_{\mathbb{R}}$ [17, Proposition VI.3.26].

Lemma 5.7 *C-tightness of a sequence $\{X^N\}$ of $\mathbb{D}_{\mathbb{R}}$-valued random elements is equivalent to the following two conditions:*

C1. The sequence of random variables $\{\|X^N\|_T\}$ is tight for every fixed $T < \infty$;

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

$$N \geq N_0 \text{ implies } \mathbb{P}(w_T(X^N, \theta) > \eta) < \varepsilon, \tag{5.22}$$

where

$$w_T(f, \theta) := \sup_{0 \leq s < u \leq s+\theta \leq T} |f(u) - f(s)|.$$

Lemma 5.8 *The sequence $\bar{\Upsilon}^N \doteq (\bar{\alpha}^N, \bar{\mu}^N, \bar{\rho}^N, \bar{e}^N), N \in \mathbb{N}$, is relatively compact in $\mathbb{D}_{\mathcal{M}} \times \mathbb{D}_{\mathbb{R}}^3$ and each of the components above is C-tight. Moreover, $\bar{e}^N \Rightarrow 0$.*

Proof: By [17, Prop. VI 1.17], to establish the first assertion of the lemma, it suffices to establish the C-tightness of each of the sequences $\{\bar{\alpha}^N\}$, $\{\bar{\mu}^N\}$, $\{\bar{\rho}^N\}$, and $\{\bar{e}^N\}$. The C-tightness of $\{\bar{\alpha}^N\}$ and $\{\bar{\mu}^N\}$ is a direct consequence of Assumption 5.2.

To show C-tightness of $\{\bar{\rho}^N\}$, fix $T < \infty$, and for $t \in [0, T-\delta]$, apply (5.10), first with $x = t$ and then with (x,t) replaced by $(t+\delta, t+\delta)$, and use (5.17) and the fact that $\beta^N, \alpha^N, \rho^N \in \mathcal{D}^{\uparrow}_{\mathcal{M}}$ to obtain

$$0 \leq \bar{\rho}^N(t+\delta) - \bar{\rho}^N(t) \leq \bar{\alpha}^N_{t+\delta}[0, t+\delta] - \bar{\alpha}^N_t[0,t] \leq \bar{\alpha}^N_t(t, t+\delta] + w_T(\bar{\alpha}^N[0,\infty), \delta). \quad (5.23)$$

Denoting by $F_{\bar{\alpha}^N_T}$ the map $x \mapsto \bar{\alpha}^N_T[0,x]$, this implies

$$w_T(\bar{\rho}^N, \delta) \leq w_T(F_{\bar{\alpha}^N_T}, \delta) + w_T(\bar{\alpha}^N[0,\infty), \delta). \quad (5.24)$$

Assumption 5.2(i) implies that both $\{\bar{\alpha}^N[0,\infty)\}$ and $\{F_{\bar{\alpha}^N_T}\}$ are C-tight, and so by Lemma 5.7, conditions C1 and C2 hold with $X^N = \bar{\alpha}^N[0,\infty)$ and $X^N = F_{\bar{\alpha}^N_T}$, $N \in \mathbb{N}$. The bound (5.24) then shows that conditions C1 and C2 of Lemma 5.7 also hold with $X^N = \bar{\rho}^N$, and so another application of Lemma 5.7 shows that $\{\bar{\rho}^N\}$ is C-tight.

Finally, we show that $\bar{e}^N \Rightarrow 0$. Due to (5.7), (5.9) and the fact that B^N takes values in $\{0,1\}$, it suffices to show that $N^{-1}(S(T^N(t)) - T^N(t)) \Rightarrow 0$. By (5.2), for fixed t,

$$N^{-1}|S(T^N(t)) - T^N(t)| \leq N^{-1} \sup_{u \in [0, \mu^N(t)]} |S(u) - u| = \sup_{u \in [0, \bar{\mu}^N(t)]} \frac{|S(Nu) - Nu|}{N} \Rightarrow 0,$$

using the functional law of large numbers for renewal processes and Assumption 5.2(2). This shows $\bar{e}^N \Rightarrow 0$. □

5.1.4 Proof of the Fluid Limit Theorem

This section is devoted to the proof of Theorem 5.4. By Lemma 5.8, the sequence $\{\bar{\Upsilon}^N\}$ is tight and $\bar{e}^N \Rightarrow 0$. Fix a convergent subsequence of the sequence $\{\bar{\Upsilon}^N\}$ relabel it as $\{\bar{\Upsilon}^N\}$, and denote the limit by $\Upsilon \doteq (\alpha, \mu, \rho, 0)$, and note that it takes values in $\mathbb{C}_{\mathcal{M}} \times \mathbb{C}^3_{\mathbb{R}}$ by Lemma 5.8. Since the components of (α, ρ, μ) are continuous and $\bar{e}^N \Rightarrow 0$, it follows that $(\bar{\alpha}^N, \bar{\mu}^N + \bar{\rho}^N + \bar{e}^N)$ converges in distribution to $(\alpha, \mu + \rho)$. Now, by (5.20) and the fact that the MVSM is preserved under scaling (which is easily deduced from Definition 2.9), we have $(\bar{\xi}^N, \bar{\beta}^N, \bar{\iota}^N) = \Theta(\bar{\alpha}^N, \bar{\mu}^N + \bar{\rho}^N + \bar{e}^N)$. By the continuity property of Θ established in Lemma 2.10 and the continuous mapping theorem, we then see that $(\bar{\xi}^N, \bar{\beta}^N, \bar{\iota}^N)$ converges in distribution to $(\xi, \beta, \iota) := \Theta(\alpha, \mu + \rho)$, and thus, (4.2)(i) holds.

To complete the proof of Theorem 5.4, it suffices to show that almost surely, $(\alpha, \xi, \beta, \mu, \rho, \iota)$ satisfy (4.2)(ii)–(iii). This suffices to prove Theorem 5.4 because Assumption 5.2 ensures that (α, μ) satisfy Assumption 4.5, and hence, Theorem 4.10 shows that equations (4.2)(i)–(iii) uniquely characterize the fluid model. To prove (4.2)(ii), note that Proposition 2.8 and (4.2)(i) show that (ξ, β, ι) takes values in $\mathbb{C}_{\mathcal{M}_0} \times \mathbb{C}^{\uparrow}_{\mathcal{M}_0} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$. In particular, ξ is continuous in t and each ξ_t has a continuous cumulative distribution, and hence, the convergence $\bar{\xi}^N \Rightarrow \bar{\xi}$ implies that $\bar{\xi}^N_t[0,t] \Rightarrow \xi_t[0,t]$. By (5.17), this gives $\xi_t[0,t] = 0$ for every $t \geq 0$.

It only remains to prove (4.2)(iii). We invoke Skorohod's representation theorem, by which we may assume without loss of generality that $(\alpha^N, \bar{\mu}^N, \bar{\rho}^N, \bar{e}^N) \to (\alpha, \mu, \rho, 0)$ and hence, that $(\bar{\xi}^N, \bar{\beta}^N, \iota^N) \to (\xi, \beta, \iota)$, almost surely. Note that the relation (4.2)(iii) does not follow directly from the convergence of $\bar{\xi}^N$ to ξ because the convergence of measures does not imply convergence of the infimum of their supports. We need to show that, with $\sigma(t) := \min \operatorname{supp}[\xi_t]$

and $T > 0$ fixed, one has $\int_{[0,T]} \mathbb{I}_{\{\sigma(t)>t\}} d\rho(t) = 0$ almost surely. Equivalently, by Fatou's lemma, we need to show that for every $\delta > 0$, the event

$$E_0 := \left\{ \int_{[0,T]} \mathbb{I}_{\{\sigma(t)>t+\delta\}} d\rho(t) > 0 \right\} \tag{5.25}$$

has zero probability. Let m, ν and $\nu_s, s \geq 0$, be as in Assumption 4.5, and recall that m is locally bounded away from zero. We fix $T < \infty$ and $\delta \in (0, \delta_0)$ where $\delta_0 \in (0, 1)$ is chosen to satisfy

$$\nu_s[0, 2\delta_0] < m(s) \text{ for all } s \in [0, T+1]. \tag{5.26}$$

The argument provided below is closely related to the one provided in the proof of Proposition 4.12 to show property (4.14). One would like to argue that a similar property must hold on the event E_0 of (5.25). However, since the subsequential limit (specifically, ρ and ξ) is not a priori known to be a.s. deterministic, measurability considerations must be taken into account to adapt the idea from the deterministic setting of Proposition 4.12. In particular, one must allow for the variable t appearing in (4.14) to be a random variable. The following lemma allows us to deal with this.

Lemma 5.9 *There exists a $[0, T) \cup \{\infty\}$-valued random variable τ, such that*

$$\mathbb{P}(E_0) = \mathbb{P}(E_1 \cap E_2), \tag{5.27}$$

where

$$E_1 := \{\tau < T,\ \sigma(\tau) > \tau + \delta\}, \qquad E_2 := \{\rho(\tau + \varepsilon) > \rho(\tau) \text{ for all } \varepsilon > 0\}. \tag{5.28}$$

The proof of Lemma 5.9 is relegated to Appendix B. We proceed with the proof of the theorem. To show that $\mathbb{P}(E_0) = 0$, we will argue that, given any random variable τ taking values in $[0, T) \cup \{\infty\}$, E_3 holds almost surely on E_1, that is, $\mathbb{P}(E_1 \cap E_3^c) = 0$, where

$$E_3 := \{\omega \in \Omega : \text{there exists } \varepsilon = \varepsilon(\omega) > 0 \text{ such that } \rho(\tau + \varepsilon) = \rho(\tau)\}. \tag{5.29}$$

Since $\{\tau < T\} \cap E_2 = \{\tau < T\} \cap E_3^c$, the result will then follow from (5.27).

Towards this end we fix a random variable τ as in Lemma 5.9. As we justify below, given any $0 \leq a < b$, the balance equation for jobs with deadlines in $(a, b]$ gives

$$\bar{\rho}^N(b) - \bar{\rho}^N(a) + \bar{\beta}_b^{s,N}(a, b] - \bar{\beta}_a^{s,N}(a, b] = \bar{\alpha}_b^N(a, b] - \bar{\alpha}_a^N(a, b] + \bar{\xi}_a^N(a, b]. \tag{5.30}$$

This relation can be obtained from (5.10) by substituting the four choices (a, a), (a, b), (b, a) and (b, b) for (t, x), and using the fact that $\bar{\xi}_b^N(a, b] = 0$ due to (5.17). Let $\delta_K = K^{-1}\delta$ for some $K \in \mathbb{N}$ and let I_k, $k = 1, \ldots, K$, denote the following partition of $(\tau, \tau + \delta]$:

$$I_k = (t_{k-1}, t_k], \qquad t_k := \tau + k\delta_K, \quad k = 1, \ldots, K.$$

By (5.30), for each N,

$$\bar{\rho}^N(\tau + \delta) - \bar{\rho}^N(\tau) = \sum_{k=1}^{K} (\bar{\rho}^N(t_k) - \bar{\rho}^N(t_{k-1})) \leq C_{N,K} + D_{N,K},$$

where

$$C_{N,K} := \sum_{k=1}^{K} \bar{\xi}^N_{t_{k-1}}(I_k), \quad \text{and} \quad D_{N,K} := \sum_{k=1}^{K} [\bar{\alpha}^N_{t_k}(I_k) - \bar{\alpha}^N_{t_{k-1}}(I_k)].$$

Now, note that

$$C_{N,K} \leq K \max_{s\in[\tau,\tau+\delta]} \bar{\xi}^N_s(\tau, \tau+\delta].$$

Now, fix K and send $N \to \infty$. Recall that we have the almost sure convergence $\bar{\xi}^N \to \xi$, as $N \to \infty$, and that $\xi \in \mathbb{C}_{\mathcal{M}_0}$. In particular, every ξ_t has a continuous distribution. Therefore, we have

$$\sup_{s\in[0,T]} d_{\mathcal{L}}(\bar{\xi}^N_s, \xi_s) \to 0, \quad \text{as} \quad N \to \infty, \quad \text{a.s.}$$

This implies that

$$\sup_{s\in[0,T]} \sup_{a\in\mathbb{R}_+} |\bar{\xi}^N_s(a,\infty) - \xi_s(a,\infty)| \to 0, \quad \text{as} \quad N \to \infty, \quad \text{a.s.}$$

On the event E_1, it must be that $\xi_\tau[\tau, \tau+\delta] = 0$, which when combined with the relation $\xi_\tau[0,\tau] = 0$ that follows from property (4.2)(ii), implies $\xi_\tau[0, \tau+\delta] = 0$. Thus, it follows that $\mathbb{I}_{E_1}\bar{\xi}^N_\tau[\tau, \tau+\delta] \to 0$ as $N \to \infty$. Now, since (4.2)(i) holds, that is, $(\xi, \beta, \iota) = \Theta(\alpha, \mu+\rho)$, (2.8) of Lemma 2.7 and the shift property of Γ_1 stated in Lemma (2.3)(1) imply that for every $t, z \geq 0$, $\xi_{t+\cdot}[0, z] = \Gamma_1(\psi^{z,t})$, where for $s \geq 0$,

$$\psi^{z,t}(s) := \xi_t[0, z] + \alpha^t_s[0, z] - \mu^t(s) - \rho^t(s).$$

Here (as in Lemma 2.3) we have used the notation $\alpha^t_s[0, z] := \hat{\alpha}_{t+s}[0, z] - \hat{\alpha}_t[0, z]$, $\mu^t(s) = \mu(t+s) - \mu(t)$, $\rho^t(s) = \rho(t+s) - \rho(t)$. Setting $z = t+\delta$, we see from (4.3) of Assumption 4.5 that for $s \geq 0$,

$$\alpha^t_s[0, t+\delta] - \mu^t(s) = \int_t^{t+s} \mathbb{I}_{\{t+\delta\geq u\}}\nu_u[0, t+\delta-u]du - \int_t^{t+s} m(u)du,$$

which is non-increasing for $s \in [0, \delta_0]$ and $t \in [0, T-\delta_0]$ due to (5.26). For each ω, applying the above with $t = \tau = \tau(\omega)$, and using the fact that $\xi_\tau[0, \tau+\delta] = 0$ on E_1, we see that $\xi_t[0, \tau+\delta] = 0$ for all $t \in [\tau, \tau+\delta]$ on E_1. As a result, for K fixed, $\lim_{N\to\infty} \mathbb{I}_{E_1} C_{N,K} = 0$ almost surely.

Next, for K fixed, it follows from Assumption 5.2(1) that $D_{N,K}$ converges almost surely, as $N \to \infty$, to

$$\begin{aligned} D_K &:= \sum_{k=1}^{K} \Big(\int_{t_{k-1}}^{t_k} \mathbb{I}_{\{t_k\geq u\}}\nu_u[0, t_k - u]du - \int_{t_{k-1}}^{t_k} \mathbb{I}_{\{t_{k-1}\geq u\}}\nu_u[0, t_{k-1} - u]du \Big) \\ &= \sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} \mathbb{I}_{\{t_k\geq u\}}\nu_u[0, t_k - u]du \\ &\leq (T+\delta) \sup_{s\in[0,T]} \nu_s[0, \delta_K]. \end{aligned}$$

By the assumption on $\nu_s, s \geq 0$, in Assumption 4.5, $D_K \to 0$ almost surely as $K \to \infty$.

Combining the estimates on $C_{N,K}$ and $D_{N,K}$, it follows that, as $N \to \infty$, $\mathbb{I}_{E_1}(\bar{\rho}^N(\tau+\delta) - \bar{\rho}^N(\tau)) \to 0$ almost surely. Now, since $\rho(t), t \geq 0$, is a continuous process, the convergence $\bar{\rho}^N \to \rho$ holds in the u.o.c. topology. As a result, $\mathbb{I}_{E_1}(\rho(\tau+\delta) - \rho(\tau)) = 0$ almost surely. This shows (5.29), which in turn establishes (4.2)(iii) and hence, completes the proof. □

5.2 Convergence Results for Policies that use Job Size Priority

We now turn to the SJF and SRPT policies. In Section 5.2.1 we introduce the primitive processes that are common to both policies, and the assumptions that we make on them. Then, in Sections 5.2.2 and 5.2.3 we introduce the state processes for the stochastic model and the associated dynamic equations for the SJF and SRPT policies, respectively, and state and prove the fluid limit convergence results, Theorems 5.13 and 5.16.

5.2.1 Common Primitive Processes and Auxiliary Processes

As before, we fix a scaling parameter N. To describe the dynamics in the N-system for both the SJF and SRPT policies, we introduce measure-valued processes that keep track of the job sizes, in addition to those that record the number of jobs. We will say that a measure $\nu \in \mathcal{M}$ is *discrete* if it is a finite sum $\sum c_i \delta_{x_i}$ of point masses, where x_i and c_i are non-negative. The weight that ν has at $x \in \mathbb{R}_+$ is, by definition, $\nu(\{x\})$.

The job-size (resp., job-count) arrival process, $\hat{\alpha}^{w,N}$ (resp., $\hat{\alpha}^{n,N}$) has sample paths in $\mathbb{D}^{\uparrow}_{\mathcal{M}}$. Here, w is a mnemonic for *work* and n for *number*, where work and job size is measured in terms of the time required to process the job at a unit service rate. For $t \geq 0$, $\hat{\alpha}^{w,N}_t$ and $\hat{\alpha}^{n,N}_t$ are discrete, and given by

$$\hat{\alpha}^{w,N}_t(dx) = \sum_{i=1}^{\infty} \mathbb{I}_{\{t \geq \tau_i\}} W_i \delta_{W_i}(dx), \qquad \hat{\alpha}^{n,N}_t(dx) = \sum_{i=1}^{\infty} \mathbb{I}_{\{t \geq \tau_i\}} \delta_{W_i}(dx), \tag{5.31}$$

where $\{\tau_i\} = \{\tau_i^N\}$ is the sequence of $\mathbb{R}_+$-valued random variables representing the arrival times of jobs into the system and $\{W_i\} = \{W_i^N\}$ is the corresponding sequence of $(0,\infty)$-valued random variables representing job sizes. Thus, $\hat{\alpha}^{w,N}_t[0,x]$ represents the amount of work that arrived in the interval $[0,t]$ due to jobs with size less than or equal to x, and $\hat{\alpha}^{n,N}_t[0,x]$ denotes the number of such jobs. Note that $\hat{\alpha}^{n,N}_t$ and $\hat{\alpha}^{w,N}_t$ can be recovered from each other via the relations

$$\hat{\alpha}^{w,N}_t[0,x] = \int_{[0,x]} y \hat{\alpha}^{n,N}_t(dy), \tag{5.32}$$

and

$$\hat{\alpha}^{n,N}_t[0,x] = \hat{\alpha}^{n,N}_t(0,x] = \int_{(0,x]} y^{-1} \hat{\alpha}^{w,N}_t(dy). \tag{5.33}$$

Also, let $m^N(t)$ denote the available rate of service at time t, and let $\mu^N(t) := \int_0^t m^N(s)ds$.

As in the case of EDF, we will also introduce some auxiliary processes that are useful for the analysis. Let B^N be a right-continuous process defined by

$$B^N(t) := \begin{cases} 1 & \text{if the server is busy at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

The processes defined by $T^N(t) := \int_0^t m^N(s)B^N(s)ds$ and $\iota^N(t) := \int_0^t m^N(s)(1-B^N(s))ds$, respectively, then represent the work done by the server and the lost work. Note that we then have the relation

$$\mu^N = T^N + \iota^N. \tag{5.34}$$

We will also introduce a state process $\xi^{w,N}$ that represents the workload in the system, whose precise definition we defer to Sections 5.2.2 and 5.2.3, since it is defined slightly differently for the SJF and SRPT policies. The value of the state just prior to zero will be denoted by $\xi^{w,N}_{0-}$, and for $N \in \mathbb{N}$, we set

$$\alpha^{w,N}_t[0,x] := \xi^{w,N}_{0-}[0,x] + \hat{\alpha}^{w,N}_t[0,x], \qquad x,t \geq 0. \tag{5.35}$$

and

$$\alpha^{n,N}_t[0,x] = \alpha^{n,N}_t(0,x] := \int_{(0,x]} y^{-1}\alpha^{w,N}_t(dy), \qquad x > 0, t \geq 0. \tag{5.36}$$

We will make the following assumptions on the primitives. Let $\bar{\alpha}^{w,N}$, $\bar{\alpha}^{n,N}$, $\bar{\mu}^N$ be the corresponding fluid-scaled quantities, defined analogously to (5.21).

Assumption 5.10 *The following two properties hold:*

(1) There exists some non-random $(\alpha^w, \mu) \in \mathbb{C}^{\uparrow}_{\mathcal{M}_0} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$ *such that*

$$(\bar{\alpha}^{w,N}, \bar{\mu}^N) \Rightarrow (\alpha^w, \mu);$$

(2) For each $0 < T < \infty$ *one has* $\int y^{-1}\alpha^w_T(dy) < \infty$ *and that the following uniform integrability condition is satisfied:*

$$\lim_{r\to\infty} \sup_N \mathbb{P}\Big(\int_{(0,\infty)} y^{-1}\mathbb{I}_{\{y^{-1}>r\}}\bar{\alpha}^{w,N}_T(dy) > \varepsilon\Big) = 0, \quad \textit{for every } \varepsilon > 0. \tag{5.37}$$

Remark 5.11 *Assumption 5.10 together with (5.36) will imply also the weak convergence of* $\bar{\alpha}^{n,N}$ *to a limit* α^n.

Example 5.12 Example 5.3(a)–(c) can be adapted to the present setting to identify conditions under which Assumption 5.10 holds. Again, let the initial conditions be zero. An analogue of Example 5.3(a) is as follows. Consider an arrival process $S^N_{\rm arr}$ that follows the same structure. Instead of $\{P_i\}$, consider an iid sequence $\{W_i\}$ of job sizes with common distribution ν^0, assumed to be atomless. Then $\hat{\alpha}^{w,N}_t$ has the form given on the left side of (5.31), with τ_i being the jump times of $S^N_{\rm arr}$. By the convergence of $N^{-1}S^N_{\rm arr}(t) \Rightarrow \lambda_0 t$ locally uniformly in t, and the LLN, the corresponding $\bar{\alpha}^{w,N} = N^{-1}\hat{\alpha}^{w,N}$ converge to α^w, where

$$\alpha^w_t(dy) = t\alpha_0 y\nu^0(dy).$$

We have $\int y^{-1}\alpha^w_T(dy) = T < \infty$, and so the first part of Assumption 5.10(2) holds. Next, the measure $y^{-1}\bar{\alpha}^{w,N}_T(dy)$ is given by

$$y^{-1}\bar{\alpha}^{w,N}_T(dy) = N^{-1}\sum_{i=1}^{\infty} \mathbb{I}_{\{t\geq\tau_i\}}\delta_{W_i}(dy).$$

Hence by the LLN, as $N \to \infty$,

$$\int_{(0,\infty)} y^{-1} \mathbb{I}_{\{y^{-1}>r\}} \bar{\alpha}_T^{w,N}(dy) = N^{-1} \sum_{i=1}^{\infty} \mathbb{I}_{\{t \geq \tau_i\}} \mathbb{I}_{\{W_i < r^{-1}\}} \Rightarrow \lambda_0 \nu^0[0, r^{-1}].$$

The validity of (5.37) follows from this.

To identify time-varying distributions that satisfy Assumption , Examples 5.3(b) and (c) can be adapted along the same lines.

5.2.2 Convergence results for the SJF Model

We now describe the state processes $\xi^{w,N}$ and $\beta^{w,N}$ for the SJF model, which have sample paths in $\mathbb{D}_{\mathcal{M}}$ and $\mathbb{D}_{\mathcal{M}}^{\uparrow}$, respectively. For $x, t > 0$, let $\xi_t^{w,N}[0,x]$ represent the total work associated with jobs that have sizes within $[0,x]$ and are present in the queue at time t, not counting the job that is at the server, and let $\beta_t^{w,N}[0,x]$ be the total work associated with jobs that have sizes within the interval $[0,x]$ that were sent to the server by time t. We let $\xi^{n,N}$ and $\beta^{n,N}$ denote the corresponding job count processes. The total work and job count measures just prior to zero are denoted by $\xi_{0-}^{w,N}$ and $\xi_{0-}^{n,N}$, respectively. We also introduce another auxiliary process, $J^N(t)$ which denotes the residual work of the job that is in service at time t. Each time the server becomes available, it admits into service the job with the smallest job size, where in case there are multiple such jobs, one of them is chosen according to some specified rule (the details of which are irrelevant for the scaling limit).

Recalling the definitions of $\alpha^{w,N}$, T^N and ι^N from Section 5.2.1, we see that the following equations then describe the system dynamics: for $t, x \geq 0$,

$$\xi_t^{w,N}[0,x] = \alpha_t^{w,N}[0,x] - \beta_t^{w,N}[0,x], \tag{5.38}$$

and

$$\beta_t^{w,N}[0,\infty) = T^N(t) + J^N(t) - J^N(0). \tag{5.39}$$

The last two equations, together with (5.34), then show that

$$\xi_t^{w,N}[0,x] = \alpha_t^{w,N}[0,x] - \mu^N(t) + \beta_t^{w,N}(x,\infty) + \iota^N(t) - J^N(t) + J^N(0), \tag{5.40}$$

and

$$\xi_t^{w,N}[0,\infty) = \alpha_t^{w,N}[0,\infty) - \mu^N(t) + \iota^N(t) - J^N(t) + J^N(0). \tag{5.41}$$

The conditions reflecting prioritization according to the size of job and non-idling, respectively, give the following two relations:

$$\int_{[0,\infty)} \xi_t^{w,N}[0,x] d\beta_t^{w,N}(x,\infty) = 0, \tag{5.42}$$

$$\int_{[0,\infty)} \xi_t^{w,N}[0,\infty) d\iota^N(t) = 0. \tag{5.43}$$

Relations (5.40)–(5.43) also hold for the scaled processes (such as $\bar{\alpha}^{w,N}$) defined by normalizing by N in a matter analogous to (5.21), and can be written in terms of the map Θ as follows:

$$(\bar{\xi}^{w,N}, \bar{\beta}^{w,N}, \bar{\iota}^N) = \Theta(\bar{\alpha}^{w,N}, \bar{\mu}^N + \bar{J}^N - \bar{J}^N(0)). \tag{5.44}$$

Note that this relation is much simpler than the corresponding (unscaled) equation (5.20) for the hard EDF policy. Since the map Θ has only been defined when the second argument of the map lies in $\mathbb{D}_{\mathbb{R}}^{\uparrow}$, it must be argued that the sample paths of $\bar{\mu}^N + \bar{J}^N - \bar{J}^N(0)$ lie in $\mathbb{D}_{\mathbb{R}}^{\uparrow}$. Indeed, this follows on writing $\bar{\mu}^N + \bar{J}^N - J^N(0) = (\bar{\mu}^N - \bar{T}^N) + (\bar{T}^N + \bar{J}^N - \bar{J}^N(0))$ and noticing that the first term lies in $\mathbb{D}_{\mathbb{R}}^{\uparrow}$ by the definitions of T^N and μ^N, and the second term lies in $\mathbb{D}_{\mathbb{R}}^{\uparrow}$ due to (5.39).

The processes $\xi^{n,N}$ and $\beta^{n,N}$ can be recovered from the above processes using the transformation (5.33), and consequently so can the normalized processes. In other words, we have

$$\bar{\xi}_t^{n,N}[0,x] = \int_{(0,x]} y^{-1}\bar{\xi}_t^{w,N}(dy), \qquad \bar{\beta}_t^{n,N}[0,x] = \int_{(0,x]} y^{-1}\bar{\beta}_t^{w,N}(dy). \tag{5.45}$$

Denote $\alpha_t^{n,N}[0,x] := \xi_{0-}^{n,N}[0,x] + \hat{\alpha}_t^{n,N}[0,x]$, and let $\bar{\alpha}^{n,N}$ be the corresponding scaled quantity.

We now state the convergence result for the SJF scheduling policy.

Theorem 5.13 *Suppose Assumption 5.10(1) holds. Then, as $N \to \infty$, we have*

$$(\bar{\xi}^{w,N}, \bar{\beta}^{w,N}, \bar{\iota}^N) \Rightarrow (\xi^w, \beta^w, \iota) := \Theta(\alpha^w, \mu). \tag{5.46}$$

If, in addition, Assumption 5.10(2) holds, then $(\bar{\xi}^{n,N}, \bar{\beta}^{n,N}) \Rightarrow (\xi^n, \beta^n)$, where

$$\xi_t^n[0,x] = \int_{(0,x]} y^{-1}\xi_t^w(dy), \qquad \beta_t^n[0,x] = \int_{(0,x]} y^{-1}\beta_t^w(dy). \tag{5.47}$$

Note that in the second part of the above result, thanks to the fact that the limits are deterministic, one has in fact joint convergence $(\bar{\xi}^{w,N}, \bar{\beta}^{w,N}, \bar{\iota}^N, \bar{\xi}^{n,N}, \bar{\beta}^{n,N}) \Rightarrow (\xi^w, \beta^w, \iota, \xi^n, \beta^n)$.

The proof of Theorem 5.13 will rely on the following two general results on tightness of measure-valued processes.

Lemma 5.14 *Let ζ and $\zeta^N, N \in \mathbb{N}$, be $\mathbb{D}_{\mathcal{M}}$-valued random elements defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that satisfy $\langle f, \zeta^N \rangle \Rightarrow \langle f, \zeta \rangle$ for every $f \in \mathbb{C}_b[0,\infty)$. Then $\zeta^N \Rightarrow \zeta$ if and only if the following compact containment condition is satisfied: for each $T > 0$ and $\eta > 0$ there exists a compact set $\mathcal{K}_{T,\eta} \subset \mathcal{M}$ such that*

$$\liminf_{N\to\infty} \mathbb{P}\left(\zeta_t^N \in \mathcal{K}_{T,\eta} \text{ for all } t \in [0,T]\right) > 1 - \eta. \tag{5.48}$$

Proof: Let $\mathbb{F}$ be the class of functionals F on $\mathcal{M}$ of the form $F = \langle f, \mu \rangle$, $\mu \in \mathcal{M}$, for some $f \in \mathbb{C}_b[0,\infty)$. Then clearly $\mathbb{F}$ is closed under addition and separates points (i.e., measures). Thus the lemma follows from [18, Theorem 3.1]. □

We now establish a useful lemma for verifying the compact containment condition. The proof of Lemma 5.15 is relegated to Appendix C.

Lemma 5.15 *Suppose the sequences $\{\zeta^N\}$ and $\{\tilde{\zeta}^N\}$ of, respectively, $\mathbb{D}_{\mathcal{M}}^{\uparrow}$-valued and $\mathbb{D}_{\mathcal{M}}$-valued random elements, are such that $\{\zeta^N\}$ satisfies the compact containment condition* (5.48) *and almost surely,*

$$\tilde{\zeta}_t^N(A) \le \zeta_t^N(A), \quad N \in \mathbb{N}, A \in \mathcal{B}(\mathbb{R}_+), t \ge 0. \tag{5.49}$$

Then $\{\tilde{\zeta}^N\}$ also satisfies the compact containment condition.

Proof of Theorem 5.13 1. Since $t \mapsto \alpha^w_t[0,\infty)$ is continuous and

$$\max_{t\in[0,T]} J^N(t) \le J^N(0) \vee \max_{t\in[0,T]} (\alpha^{w,N}_t - \alpha^{w,N}_{t-}), \quad T > 0,$$

it follows that $\bar{J}^N \Rightarrow 0$ as $N \to \infty$. In light of the continuity of Θ stated in Proposition 2.10(1), the first assertion then follows by an application of the continuous mapping theorem using (5.44), the limit $\bar{J}^N \Rightarrow 0$, and the assumed convergence of $(\bar{\alpha}^{w,N}, \bar{\mu}^N) \Rightarrow (\alpha^w, \mu)$, for some non-random $(\alpha^w, \mu) \in \mathbb{C}^{\uparrow}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$. Moreover, Proposition 2.8 shows that the limit (ξ^w, β^w, ι) lies in $\mathbb{C}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$.

2. We start by fixing $f \in \mathcal{C}_b[0,\infty)$, and showing that

$$\langle f, \bar{\xi}^{n,N} \rangle \Rightarrow \langle f, \xi^n \rangle, \qquad \text{and} \qquad \langle f, \bar{\beta}^{n,N} \rangle \Rightarrow \langle f, \beta^n \rangle. \tag{5.50}$$

To prove (5.50), note that we may assume, without loss of generality, that $f \ge 0$. Since α^w is assumed to be in $\mathbb{C}^{\uparrow}_{\mathcal{M}_0}$ in the second part of Theorem 5.13, it follows from Proposition 2.8 that $\xi^w, \beta^w \in \mathbb{C}_{\mathcal{M}_0}$. Hence the convergence $\bar{\xi}^{w,N} \Rightarrow \xi^w$, as $N \to \infty$, proved in part 1 of the theorem, implies $\langle f, \bar{\xi}^{w,N} \rangle \Rightarrow \langle f, \xi^w \rangle$. Thus, for any $T, r < \infty$, as $N \to \infty$,

$$A_{N,r} := \sup_{t\in[0,T]} \left| \int (y^{-1} \wedge r) f(y) \bar{\xi}^{w,N}_t(dy) - \int (y^{-1} \wedge r) f(y) \xi^w_t(dy) \right| \to 0, \quad \text{in probability.} \tag{5.51}$$

Fix $\delta > 0$ and $\varepsilon > 0$. Then we have

$$\begin{aligned} D_{N,r} &:= \sup_{t\in[0,T]} \int y^{-1} f(y) \bar{\xi}^{w,N}_t(dy) - \int (y^{-1} \wedge r) f(y) \bar{\xi}^{w,N}_t(dy) \\ &= \sup_{t\in[0,T]} \int y^{-1} f(y) \mathbb{I}_{\{y^{-1} > r\}} \bar{\xi}^{w,N}_t(dy) \\ &\le \int y^{-1} f(y) \mathbb{I}_{\{y^{-1} > r\}} \bar{\alpha}^{w,N}_T(dy), \end{aligned}$$

where the last inequality uses (5.38). Thus, using (5.37), one can select r sufficiently large so that

$$\sup_N \mathbb{P}\Big(D_{N,r} > \frac{\varepsilon}{3} \Big) < \frac{\delta}{2}.$$

Since $(\xi^w, \beta^w, \iota) = \Theta(\alpha^w, \mu)$, by property (2.6) of Θ, it follows that the measure α^w_T dominates ξ^w_T and β^w_T, in the sense that $\xi^w_T(B), \beta^w_T(B) \le \alpha^w_T(B)$ for every Borel set $B \subset \mathbb{R}_+$. Hence, the moment assumption $\int y^{-1} \alpha^w_T(dy) < \infty$ implies that the same estimate holds when α^w_T is replaced by either ξ^w_T or β^w_T, and $\alpha^{w,N}_T$ is replaced by α^w_T. Thus, by making r larger if needed, one also has

$$C_r := \sup_{t\in[0,T]} \left| \int y^{-1} f(y) \xi^w_t(dy) - \int (y^{-1} \wedge r) f(y) \xi^w_t(dy) \right| < \frac{\varepsilon}{3}.$$

For fixed r as above, let N_0 be large enough such that for all $N > N_0$, $\mathbb{P}(|A_{N,r}| > \varepsilon/3) < \delta/2$, which is possible due to (5.51). Combining the bounds on $A_{N,r}$, $D_{N,r}$ and C_r, one has

$$\limsup_N \mathbb{P}\Big(\sup_{t\in[0,T]} \Big| \int f(y) \bar{\xi}^{n,N}_t(dy) - \int f(y) \xi^n_t(dy) \Big| > \varepsilon \Big) < \delta.$$

Since δ and ε are arbitrary, we have proved (5.50). An exactly analogous proof shows that $\langle f, \bar{\beta}^{n,N} \rangle \Rightarrow \langle f, \beta^n \rangle$.

In view of Lemma 5.14, to show that $\bar{\xi}^{n,N} \Rightarrow \xi^n$ and $\bar{\beta}^{n,N} \Rightarrow \beta^n$ it only remains to show that $\{\bar{\xi}^{n,N}\}$ and $\{\bar{\beta}^{n,N}\}$ satisfy the compact containment condition. First note that (5.38) implies that for any Borel set $B \subset \mathbb{R}_+$, $\bar{\xi}^{w,N}_t(B) \le \bar{\alpha}^{w,N}_t(B)$ and $\bar{\beta}^{w,N}_t(B) \le \bar{\alpha}^{w,N}_t(B)$. Together with (5.33) and (5.45) this implies that for every Borel set $B \subset \mathbb{R}_+$,

$$\bar{\xi}^{n,N}_t(B) \le \bar{\alpha}^{n,N}_t(B), \qquad \bar{\beta}^{n,N}_t(B) \le \bar{\alpha}^{n,N}_t(B). \tag{5.52}$$

Now, Remark 5.11 and the fact that $\bar{\alpha}^{w,N} \Rightarrow \alpha^w$ imply that $\bar{\alpha}^{n,N} \Rightarrow \alpha^n$. Thus by Lemma 5.14, $\{\bar{\alpha}^{n,N}\}$ satisfies the compact containment condition. In turn, Lemma 5.15 and (5.52) together imply that $\{\bar{\xi}^{n,N}\}$ and $\{\bar{\beta}^{n,N}\}$ also satisfy the compact containment condition. Lemma 5.14 and the convergence $\langle f, \bar{\beta}^{n,N} \rangle$ and $\langle f, \bar{\xi}^{n,N} \rangle$ established above then imply that $\beta^{n,N} \Rightarrow \beta^n$ and $\bar{\xi}^{n,N} \Rightarrow \xi^n$.

Since both limits β^n and ξ^n are deterministic, to deduce joint convergence, it suffices to show that both β^n and ξ^n are members of $\mathbb{C}_{\mathcal{M}}$. To this end, we use again the fact that the measure $y^{-1}\alpha^w_T(dy)$, that is finite by assumption, dominates ξ^n_t and β^n_t for all $t \in [0,T]$. We argue that in view of this, $t \mapsto \xi^n_t$ inherits continuity from $t \mapsto \xi^w_t$. Given $g \in \mathbb{C}_b(\mathbb{R}_+)$ and $t_k \to t$, we have, for any $\varepsilon > 0$,

$$\begin{aligned} &\Big| \int y^{-1} g(y) \xi^w_{t_k}(dy) - \int y^{-1} g(y) \xi^w_t(dy) \Big| \\ &\qquad \le 2\|g\| \int_{(0,\varepsilon)} y^{-1} \alpha^w_T(dy) + \Big| \int_{[\varepsilon,\infty)} y^{-1} g(y) \xi^w_{t_k}(dy) - \int_{[\varepsilon,\infty)} y^{-1} g(y) \xi^w_t(dy) \Big|. \end{aligned}$$

The last term on the right-hand side converges to zero as $k \to \infty$, and since ε is arbitrary, so does the left-hand side. Thus, $\xi^n \in \mathbb{C}_{\mathcal{M}}$. Similarly, $\beta^n \in \mathbb{C}_{\mathcal{M}}$. This completes the proof. □

5.2.3 Convergence Results for the SRPT Model

We recall the primitive processes $\hat{\alpha}^{w,N}$, $\hat{\alpha}^{n,N}$, m^N, μ^N, B^N, T^N and ι^N introduced in Section 5.2.1. We denote the state processes for the SRPT model also by $\xi^{w,N}$ and $\beta^{w,N}$, although they are now defined somewhat differently from the SJF model. For the in-queue job size measure, $\xi^{w,N}$, under the SRPT policy, it is more convenient to work with a version that includes the job that is being served at the current time. More precisely, $\xi^{w,N}$, is a process with sample paths in $\mathbb{D}_{\mathcal{M}}$ that now records the *initial job requirements* associated with all jobs that are still in the system, i.e., that have not yet been fully served (see Remark 5.18 for our results regarding a closely related process). The process $\beta^{w,N}$, with sample paths in $\mathbb{D}^{\uparrow}_{\mathcal{M}}$, is defined to be such that $\beta^{w,N}_t[0,x]$ denotes the total work associated with jobs that by time t have departed the system, for which the *initial* job size is within $[0,x]$. The processes $\xi^{n,N}$ and $\beta^{n,N}$ denote the corresponding job counts. As in Sections 5.2.1 and 5.2.2, let $\alpha^{w,N} = \xi^{w,N}_{0-} + \hat{\alpha}^{w,N}$ and $\alpha^{n,N} = \xi^{n,N}_{0-} + \hat{\alpha}^{n,N}$, and let the quantities $\bar{\alpha}^{w,N}, \bar{\alpha}^{n,N}$ and $\bar{\mu}^N$ denote the corresponding scaled quantities as in (5.21).

We may (and will) assume, without loss of generality, that all jobs present in the system at time $t = 0$ have not been processed before (even the job that is at service at this time). Indeed, given an arbitrary initial configuration where some jobs are partially served at time zero, the system will behave in exactly the same way as under an initial configuration in which all the portions of service that were already provided are forgotten. Thus, the initial condition ξ^N_{0-}, which encodes the residual service times, will be treated as if these were the original sizes of jobs (note that we do not make any explicit distributional assumptions on either the original job sizes or these residual sizes beyond the convergence in Assumption 5.10).

We now state the main convergence result for the SRPT model.

Theorem 5.16 *Suppose Assumption 5.10 holds. Then*

$$(\bar{\xi}^{w,N}, \bar{\beta}^{w,N}, \bar{\iota}^N) \Rightarrow (\xi^w, \beta^w, \iota) := \Theta(\alpha^w, \mu), \tag{5.53}$$

and $(\bar{\xi}^{n,N}, \bar{\beta}^{n,N}) \Rightarrow (\xi^n, \beta^n)$*, where* ξ^n *and* β^n *are as defined in* (5.47)*.*

Remark 5.17 Note that, in contrast to the corresponding result for the SJF policy, namely Theorem 5.13, our proof of even the limit (5.53) for the SRPT policy requires both parts of Assumption 5.10, and not just Assumption 5.10(1).

For the proof of the theorem, and to describe the dynamics, it will be convenient to introduce some terminology to distinguish the different states of jobs. Jobs that have not departed the system are said to be *in the queue* (note that this includes the job being served). Jobs in the queue can be in one of two states: *partially served*, by which we refer to a job that is either being served at the moment or has been previously served but was preempted by another job, or *unserved*, by which we mean a job that has arrived but has not yet been served. We further distinguish partially served jobs according to whether u units of the job size have or have not been processed, where u is a given threshold. The main idea of the proof is as follows. We argue that for a suitable choice of $u = u^N$, at any given time only a small number of jobs have a size that is u or more units smaller than the initial size. On the other hand, we show that jobs in the complement set (namely partially served jobs for which less than u units of work has been processed) can be treated as unserved, since the resulting error is small due to the fact that their residual job sizes do not deviate much from their initial job sizes.

To formulate this notion, we recall that $W_i = W_i^N$ denotes the size of job i, and $\tau_i = \tau_i^N$ denotes the time of arrival into system of that job. Let $W_i^N(t)$ denote the residual job size in the N-system at time t (defined only for $t \geq \tau_i^N$). Note that by our assumption, $W_i^N(0) = W_i^N$ for all jobs i that are in the system at time 0. Then, after possibly relabeling the job sizes, we can express the process $\alpha^{w,N}$ as in (5.31) to be of the form

$$\alpha_t^{w,N}(A) = \sum_{i=1}^{\infty} \mathbb{I}_{\{t \geq \tau_i^N\}} W_i^N \delta_{W_i^N}(A), \quad A \in \mathcal{B}(\mathbb{R}_+), \tag{5.54}$$

with the convention that $\tau_i^N \leq 0$ for jobs that are initially in the system. Given a parameter $u > 0$, let $\theta_i = \theta_i^N(u) := \inf\{t \geq \tau_i : W_i^N(t) \leq W_i^N - u^N\}$, and refer to job i as *u-unserved* at time t if $\tau_i^N \leq t < \theta_i^N$. Note that this includes unserved jobs (that have already arrived by time t) and others, partially served, for which less than u units of their processing requirements have been processed prior to that time. We also say that job i is *u-served* at time t if $t \geq \theta_i^N$, in which case at least u units of its size have been processed at that time (whether it is partially served or has departed in the interval $[0, t]$). Furthermore, we say that a job i is *u-short* if its original job size satisfies $W_i^N < u$. Note that for such a job, $\theta_i^N = \infty$, and therefore it is u-unserved even at its departure time.

The parameter u to be used will depend on N. To this end we fix a sequence $u^N > 0, N \in \mathbb{N}$, with $\lim_{N\to\infty} u^N = 0$ and $\lim_{N\to\infty} Nu^N = \infty$. In what follows, we suppress N from the constant u^N (and, in particular, use the terms u-unserved and u-served for u^N-unserved and u^N-served), and also from the random variables W_i^N, τ_i^N and $\theta_i^N = \theta_i^N(u^N)$, but we retain it for all processes such as $W_i^N(\cdot)$ and $\alpha^{w,N}$ that describe the dynamics of the N-system. We now introduce a certain modified arrival process $\alpha_t^{*,N}$. Denote

$$W_i^{u,N}(t) = \max(W_i^N(t), W_i - u),$$

and

$$\alpha_t^{*,N}(A) = \sum_{i=1}^{\infty} \mathbb{I}_{\{t\geq\tau_i\}} W_i^{u,N}(t)\delta_{W_i^{u,N}(t)}(A), \quad A \in \mathcal{B}(\mathbb{R}). \tag{5.55}$$

Note that this process has sample paths in $\mathbb{D}_{\mathcal{M}}$ (in particular, for every $t \geq 0$, $\alpha^{*,N}$ is a measure), though not necessarily in $\mathbb{D}_{\mathcal{M}}^{\uparrow}$.

We now introduce the corresponding state processes. Let $I_1^N(t)$ and $I_2^N(t)$ denote the sets of u-unserved and, respectively, u-served jobs at time t. Let $\xi^{*,N}$ be a process with sample paths in $\mathbb{D}_{\mathcal{M}}$ recording the residual job sizes of u-unserved jobs, given by

$$\xi_t^{*,N}(A) = \sum_{i\in I_1^N(t)} \mathbb{I}_{\{t\geq\tau_i\}} W_i^N(t)\delta_{W_i^N(t)}(A), \quad A \in \mathcal{B}(\mathbb{R}_+). \tag{5.56}$$

Accordingly, let $\beta^{*,N}$ be a process with sample paths in $\mathbb{D}_{\mathcal{M}}^{\uparrow}$, recording work that has departed from the class of u-unserved jobs. More precisely,

$$\beta_t^{*,N}(A) = \sum_{i\in I_2^N(t)} (W_i - u)\delta_{(W_i-u)}(A), \quad A \in \mathcal{B}(\mathbb{R}_+). \tag{5.57}$$

Note that $\beta_t^{*,N}[0,x]$ is the sum of the residual job sizes at the time of becoming u-served, of u-served jobs whose residual job size at that time lies in the interval $[0,x]$. Note that u-short jobs never become members of $I_2^N(t)$ for any t, and therefore their job sizes are not recorded in $\beta^{*,N}$. Next, let $J_2^{n,N}(t)$ denote the number of all partially served u-served jobs at time t, and let $J_2^{w,N}(t)$ denote the total residual work of all such jobs. In the proof it will be argued that the error between the processes $\xi^{*,N}$ and $\xi^{w,N}$, $\beta^{*,N}$ and $\beta^{w,N}$, etc., tends to zero in the limit.

We now write down the equations satisfied by the processes $\alpha^{*,N}$, $\xi^{*,N}$ and $\beta^{*,N}$. First, note that $I_1^N(t) \cup I_2^N(t)$ is equal to the set of all jobs i for which $t \geq \tau_i$. Also note that for $i \in I_1^N(t)$, $W_i^{u,N}(t) = W_i^N(t)$, while for $i \in I_2^N(t)$, $W_i^{u,N}(t) = W_i - u$. Therefore, (5.55), (5.56) and (5.57) yield

$$\xi_t^{*,N}[0,x] = \alpha_t^{*,N}[0,x] - \beta_t^{*,N}[0,x]. \tag{5.58}$$

Next, let $r^N(t)$ denote the total amount of work done on the u-unserved jobs by time t, that is, $r^N(t) = \sum_{i\in I_1^N(t)}(W_i - W_i^N(t))$. Recall from Section 5.2.1 that $T^N(t)$ represents the total work that was processed from all jobs in the interval $[0,t]$. Then we obtain from (5.57) that $\beta_t^{*,N}[0,\infty) = \sum_{i\in I_2^N(t)} W_i - u|I_2^N(t)|$, while

$$T^N(t) = \sum_{i\in I_1^N(t)\cup I_2^N(t)} (W_i - W_i^N(t)) = r^N(t) + \sum_{i\in I_2^N(t)} (W_i - W_i^N(t)) = r^N(t) + \sum_{i\in I_2^N(t)} W_i - J_2^{w,N}(t).$$

Combining the last two identities, we obtain

$$\beta_t^{*,N}[0,\infty) = T^N(t) + J_2^{w,N}(t) - r^N(t) - u|I_2^N(t)|. \tag{5.59}$$

Thus, combining (5.59) with (5.58) and recalling that $T^N(t) = \mu^N(t) - \iota^N(t)$, we have

$$\xi_t^{*,N}[0,x] = \alpha_t^{*,N}[0,x] - \mu^N(t) + \beta_t^{*,N}(x,\infty) + \iota^N(t) - J_2^{w,N}(t) + r^N(t) + u|I_2^N(t)|, \tag{5.60}$$

and

$$\xi_t^{*,N}[0,\infty) = \alpha_t^{*,N}[0,\infty) - \mu^N(t) + \iota^N(t) - J_2^{w,N}(t) + r^N(t) + u|I_2^N(t)|. \tag{5.61}$$

It is of crucial importance that these processes also satisfy

$$\int_{[0,\infty)} \xi_t^{*,N}[0,x] d\beta_t^{*,N}(x,\infty) = 0, \tag{5.62}$$

which reflects the fact that jobs with residual sizes greater than x cannot be served unless there are no jobs in queue with residual sizes less than or equal to x, and

$$\int_{[0,\infty)} \xi_t^{*,N}[0,\infty) d\iota^N(t) = 0, \tag{5.63}$$

which captures the fact that the server cannot be idle if there is a job with positive residual work still in the queue. As before, scaled processes are denoted using the bar notation (as in $\bar\alpha^{w,N}$), and for any set-valued process $S^N(t)$, we use $||\bar S^N||_T$ to denote $\sup_{t\in[0,T]} |S^N(t)|/N$, where $|S^N(t)|$ denotes the cardinality of $S^N(t)$.

Proof of Theorem 5.16. We start with the proof of (5.53), which proceeds via the following steps. In Step 1 we show that, for fixed T, and the fixed sequence $\{u_N\}$, $\|\bar r^N\|_T \vee u^N\|\bar I_2^N\|_T \vee \|\bar J_2^{w,N}\|_T \vee \|\bar J_2^{n,N}\|_T \to 0$ in probability. In Step 2 we show $\bar\alpha^{*,N} \Rightarrow \alpha^w$. Step 3 shows tightness of the collection of processes $(\alpha^{*,N}, \mu^N, \xi^{*,N}, \beta^{*,N}, \iota^N)$. Finally, in Step 4, limits are taken in (5.60), (5.62) and (5.63) to obtain that every subsequential limit $(\alpha^w, \mu, \xi^w, \beta^w, \iota)$ of the aforementioned sequence satisfies the relation $(\xi^w, \beta^w, \iota) = \Theta(\alpha^w, \mu)$, by which the limit in probability exists. Using estimates on the error terms from Step 1, it is then shown that the same follows regarding $(\alpha^{w,N}, \mu^N, \xi^{w,N}, \beta^{w,N}, \iota^N)$.

Step 1: For fixed T, we first show that $\|\bar J_2^{w,N}\|_T \vee \|\bar J_2^{n,N}\|_T \to 0$ in probability.

To this end, note that, as a consequence of the assumed convergence $\bar\alpha^{w,N} \Rightarrow \alpha^w$ in Assumption 5.10(1) and (5.32) one has

$$\lim_{r\to\infty} \sup_N \mathbb{P}\Big(\int_{(r,\infty)} y\bar\alpha_T^{n,N}(dy) > \varepsilon\Big) = 0, \quad \text{for every } \varepsilon > 0. \tag{5.64}$$

To address the convergence of $\bar{J}_2^{w,N}$, we shall show that for any $\varepsilon > 0$ and $\eta > 0$, one has

$$\mathbb{P}(\|\bar{J}_2^{w,N}\|_T > \eta) \leq \varepsilon, \qquad \text{for all large } N. \tag{5.65}$$

Let $\varepsilon > 0$ and $\eta > 0$ be given. By (5.64), there exists ℓ so large that

$$\sup_N \mathbb{P}\Big(\int_{(\ell,\infty)} y\bar{\alpha}_T^{n,N}(dy) > \frac{\eta}{2} \Big) \leq \varepsilon. \tag{5.66}$$

Fix such ℓ, and assume without loss of generality that $\ell > 1$. Consider the Nth system on the time interval $[0, T]$. For this argument only, let the jobs be labeled according to the order of their first admittance into service. Namely, for $i \in \mathbb{N}$, let the term 'job i' refer to the ith job to be admitted into service for the first time. For $i, j \in \mathbb{N}$, let the notation $i < j$ stand for the order thus defined. Let σ_i denote the time when job i is first admitted into service (again, the dependence on N is suppressed). Thus, σ_i is increasing with i.

Next, given $t \in [0, T]$, let $I^N(t)$ denote the collection of jobs that are u-served and partially served at time t, except the one that is being served at that time (if such a job exists). Note that the cardinality of this set is, by definition, $(J_2^{n,N}(t) - 1) \vee 0$. Then each job in $I^N(t)$ has been served and preempted prior to time t. Moreover, for each $i \in I^N(t)$, the residual work satisfies $W_i^N(t) \leq W_i - u$. Suppose $i, j \in I^N(t)$ with $j > i$. Then i and j are both partially served, and j was first admitted into service later than i was. Due to the SRPT policy, this implies that at the time σ_j that job j first receives service, the size of job j is less than that of job i, or equivalently, $W_j \leq W_i^N(\sigma_j)$. Moreover, it is impossible for job i to be processed during the time interval $[\sigma_j, t]$, because job j has not yet departed at time t because, by assumption, $j \in I^N(t)$. Thus, $W_i^N(\sigma_j) = W_i^N(t)$. Since $i \in I^N(t)$ implies $W_i^N(t) \leq W_i - u$, this means that

$$W_j \leq W_i - u, \quad \text{ whenever } i, j \in I^N(t),\ j > i.$$

Let $\hat{I}^N(t)$ denote the collection of members i of $I^N(t)$ with $W_i \leq \ell$. As a consequence of the above display, if $\hat{I}^N(t)$ is nonempty and if i_t and j_t denote minimal and, respectively, maximal members of $\hat{I}^N(t)$, then

$$0 \leq W_{j_t} \leq W_{i_t} - u(|\hat{I}^N(t)| - 1) \leq \ell - u|\hat{I}^N(t)| + u. \tag{5.67}$$

Now recall that $\alpha^{n,N} = \xi_{0-}^{n,N} + \hat{\alpha}^{n,N}$ where $\hat{\alpha}^{n,N}$ is given by (5.31) and satisfies(5.33). Hence, $J_2^{n,N}(t) \leq |I^N(t)| + 1 \leq |\hat{I}^N(t)| + \alpha_T^{n,N}[\ell, \infty) + 1$. Then, in view of (5.66) and (5.67) and the fact that $\ell > 1$, on an event whose probability is at least $1 - \varepsilon$, for any $\eta > 0$, we can bound

$$\sup_{t\in[0,T]} \bar{J}_2^{n,N}(t) \leq \frac{\ell + 2u}{uN} + \frac{\eta}{2}, \qquad \sup_{t\in[0,T]} \bar{J}_2^{w,N}(t) \leq \frac{\ell(\ell + 2u)}{uN} + \frac{\eta}{2},$$

for all large enough N. Since $uN = u^N N \to \infty$, the above two expressions are bounded by η for all sufficiently large N. Since $\eta > 0$ is arbitrary, we obtain the asserted convergence.

Next, we show that $\|\bar{r}^N\|_T \vee u\|\bar{I}_2^N\|_T \to 0$ in probability. By definition, the server has processed a portion of at most u units of work for each job in $I_1^N(t)$. Therefore, we have

$$\|\bar{r}^N\|_T \leq u\|\bar{I}_1^N\|_T \leq u\bar{\alpha}_T^{n,N}[0, \infty).$$

Similarly, note that $u\|\bar{I}_2^N\|_T \le u\bar{\alpha}_T^{n,N}[0,\infty)$. Note that Remark 5.11 and the fact that $\bar{\alpha}^{w,N} \Rightarrow \alpha^w$ imply that $\bar{\alpha}^{n,N} \Rightarrow \alpha^n$. Hence, recalling that we assume $u = u^N \to 0$, we have $\|\bar{r}^N\|_T \vee \|u\bar{I}_2^N\|_T \to 0$ in probability.

Step 2: We show that $\bar{\alpha}^{*,N} \Rightarrow \alpha^w$.

This is basically a consequence of the fact, which we will establish below, that one has $\sup_{t\in[0,T]} d_{\mathcal{L}}(\bar{\alpha}_t^{*,N}, \bar{\alpha}_t^{w,N}) \to 0$ in probability. Since $\bar{\alpha}_t^{*,N}[x,\infty)$ is dominated by $\bar{\alpha}_T^{w,N}[x,\infty)$ for all $x \ge 0$ and $t \in [0,T]$, in view of (5.66) and (1.2), it is enough to show that for every $\ell > 0$,

$$\sup_{t\in[0,T]} \sup_{x\in[0,\ell]} |\bar{\alpha}_t^{*,N}[0,x] - \bar{\alpha}_t^{w,N}[0,x]| \to 0, \quad \text{in probability,} \tag{5.68}$$

Given $t \in [0,T]$, $x \in [0,\ell]$, and $\ell < \infty$,

$$|\langle \mathbb{I}_{[0,x]}, \bar{\alpha}_t^{*,N}\rangle - \langle \mathbb{I}_{[0,x]}, \bar{\alpha}_t^{w,N}\rangle| = \vartheta_1^N(t) + \vartheta_2^N(t), \tag{5.69}$$

where

$$\begin{aligned} \vartheta_1^N(t) &= N^{-1}\sum_i \mathbb{I}_{\{t\ge\tau_i\}}\mathbb{I}_{\{W_i\le x\}}[W_i - W_i^{u,N}(t)] \\ &\le N^{-1}\sum_i \mathbb{I}_{\{t\ge\tau_i\}}\mathbb{I}_{\{W_i\le \ell\}}u \\ &\le u\,\alpha_T^{n,N}[0,\infty). \end{aligned} \tag{5.70}$$

and

$$\begin{aligned} \vartheta_2^N(t) &= N^{-1}\sum_i \mathbb{I}_{\{t\ge\tau_i\}}\mathbb{I}_{\{W_i^{u,N}(t)\le x<W_i\}}W_i^{u,N}(t) \\ &\le N^{-1}\sum_i \mathbb{I}_{\{t\ge\tau_i\}}\mathbb{I}_{\{W_i-u\le x<W_i\}}W_i \\ &\le N^{-1}\sum_i \mathbb{I}_{\{t\ge\tau_i\}}\mathbb{I}_{\{x<W_i\le x+u\}}W_i \\ &\le \sup_{x\ge 0} \bar{\alpha}_T^{w,N}(x,x+u]. \end{aligned} \tag{5.71}$$

Since $\alpha^w \in \mathbb{C}^{\uparrow}_{\mathcal{M}_0}$, $\sup_{x\ge0}\alpha_T^w(x,x+u] \to 0$ as $u = u^N \to 0$. Therefore by Assumption 5.10(1) on $\bar{\alpha}^{w,N}$ and (1.2) we note that $\vartheta_2^N \to 0$ uniformly in $t \in [0,T]$. On the other hand, using Remark 5.11 together with fact that $u = u^N \to 0$ as $N \to \infty$, we note that $\vartheta_1^N(t) \to 0$ as $u^N \to 0$ uniformly in $t \in [0,T]$. Thus, using (5.70)-(5.71) in (5.69), we obtain that the right-hand side in (5.69) converges to zero in probability uniformly in $x \le \ell$. Thus, we have established (5.68).

Step 3: We now establish the C-tightness of $(\bar{\alpha}^{*,N}, \bar{\mu}^N, \bar{\xi}^{*,N}, \bar{\beta}^{*,N}, \bar{\iota}^N)$.

From Step 2, Assumption 5.10 and (5.33) it is clear that $\{(\bar{\alpha}^{*,N}, \bar{\mu}^N, \bar{\iota}^N), N \ge 1\}$ is tight. Next, we note that for all $x \ge 0$,

$$\sup_{t\in[0,T]} \bar{\xi}_t^{*,N}[x,\infty) \vee \bar{\beta}_t^{*,N}[x,\infty) \le \bar{\alpha}_T^{w,N}[x,\infty).$$

Thus $\{(\bar{\xi}^{*,N}, \bar{\beta}^{*,N}), N \geq 1\}$ satisfies the compact containment condition stated in Lemma 5.14. Again for any $0 \leq s < t \leq T$, we get from (5.56) that

$$\sup_{x\in\mathbb{R}_+} |\bar{\xi}_t^{*,N}[0,x] - \bar{\xi}_s^{*,N}[0,x]| \leq (\bar{\alpha}_t^{w,N}(\mathbb{R}_+) - \bar{\alpha}_s^{w,N}(\mathbb{R}_+)) + (\bar{\mu}_t^N - \bar{\mu}_s^N).$$

Thus by our assumption on $(\bar{\alpha}^{w,N}, \bar{\mu}^N)$ and (1.2) we see that the oscillation of $\bar{\xi}^{*,N}$ with respect to $d_{\mathcal{L}}$ tends to zero in probability. A similar fact also holds for $\bar{\beta}^{*,N}$ due to the relation (5.58). This establishes tightness of $(\bar{\xi}^{*,N}, \bar{\beta}^{*,N})$ [11, Corollary 3.7.4]. Moreover, it is readily seen that any sub-sequential limit of $(\bar{\alpha}^{*,N}, \bar{\mu}^N, \bar{\xi}^{*,N}, \bar{\beta}^{*,N}, \bar{\iota}^N)$ is continuous in the variable t [11, Theorem 3.10.2].

Step 4: Now we characterize the limits of $\Sigma^{*,N} := (\bar{\alpha}^{*,N}, \bar{\mu}^N, \bar{\xi}^{*,N}, \bar{\beta}^{*,N}, \bar{\iota}^N)$, and in turn, of $\Sigma^{w,N} := (\bar{\alpha}^{w,N}, \bar{\mu}^N, \bar{\xi}^{w,N}, \bar{\beta}^{w,N}, \bar{\iota}^N)$.

Given any subsequence of $\Sigma^{*,N}$ which converges, and denoting by $\Sigma = (\alpha^w, \mu, \xi^w, \beta^w, \iota)$ its limit in distribution, we take limits in equations (5.59), (5.60), (5.62) and (5.63). Note that the sample paths of Σ are in $\mathbb{C}^{\uparrow}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}} \times \mathbb{C}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathcal{M}} \times \mathbb{C}^{\uparrow}_{\mathbb{R}}$ due to Step 3. Since for any $\delta > 0$, we have

$$\sup_{t\in[0,T]} \sup_{x\geq 0} \xi_t^{*,N}[x, x+\delta] \leq \sup_{x\geq 0} \alpha_T^{w,N}[x, x+\delta+u],$$

and $u = u^N \to 0$, it follows from Assumption 5.10(1) that $\xi^w \in \mathbb{C}_{\mathcal{M}_0}$. Using (5.58), we also have $\beta^w \in \mathbb{C}^{\uparrow}_{\mathcal{M}_0}$. Due to this property and the fact that $\iota \in \mathbb{C}^{\uparrow}_{\mathbb{R}}$, relations (5.62) and (5.63) are preserved under the limit. Hence, using the estimates from Step 1 in (5.59) and (5.60), it follows that Σ satisfies the four hypotheses of Definition 2.5. As a result, $(\xi^w, \beta^w, \iota) = \Theta(\alpha^w, \mu)$. Since this holds for any subsequential limit, we conclude that $\Sigma^{*,N}$ converges in probability to Σ.

Next we obtain the limit of $(\bar{\xi}^{w,N}, \bar{\beta}^{w,N})$ by comparing these processes to $(\bar{\xi}^{*,N}, \bar{\beta}^{*,N})$. Given a test function $g \in C_b(\mathbb{R})$, and given $t \in [0,T]$, it follows from (5.56) that

$$\langle g, \bar{\xi}_t^{w,N}\rangle - \langle g, \bar{\xi}_t^{*,N}\rangle = \frac{1}{N}\sum_{i=1}^{\infty} \mathbb{I}_{\{t\geq\tau_i\}}\mathbb{I}_{\{W_i^N(t)>0\}} W_i g(W_i) - \frac{1}{N}\sum_{i\in I_1^N(t)} \mathbb{I}_{\{t\geq\tau_i\}} W_i^N(t) g(W_i^N(t))).$$

For a u-unserved job i, $W_i - W_i^N(t) \leq u$, provided $t \geq \tau_i$. Hence, it follows that

$$\begin{aligned} |\langle g, \bar{\xi}_t^{w,N}\rangle - \langle g, \bar{\xi}_t^{*,N}\rangle| &\leq Osc_u(g)\bar{\alpha}_T^{w,N}[0,\infty) + \bar{r}^N\|g\|_\infty + \frac{2\|g\|_\infty}{N}\sum_{i\in I_2^N(t)} \mathbb{I}_{\{t\geq\tau_i\}}\mathbb{I}_{\{W_i^N(t)>0\}} W_i \\ &\leq Osc_u(g)\bar{\alpha}_T^{w,N}[0,\infty) + \bar{r}^N\|g\|_\infty + 2\|g\|_\infty \left\{\ell \bar{J}_2^{n,N}(t) + \bar{\alpha}_t^{w,N}[\ell,\infty)\right\}, \end{aligned}$$

for any ℓ. Sending first $N \to \infty$ and then $\ell \to \infty$, shows the convergence in probability to zero of the left-hand side, uniformly in $t \in [0,T]$. A similar estimate on $|\langle g, \bar{\beta}_t^{w,N}\rangle - \langle g, \bar{\beta}_t^{*,N}\rangle|$ follows by appealing to (5.58). Thus, the convergence of $(\xi^{w,N}, \beta^{w,N})$ follows from that of $(\xi^{*,N}, \beta^{*,N})$. The proof of part 1 is now complete.

Finally, based on part 1, the proof of part 2 of the theorem follows along the lines of the proof of part 2 of Theorem 5.13. □

Remark 5.18 It is natural to associate a measure $\xi^{w,N} \in \mathbb{D}_{\mathcal{M}}$ with the queue length process defined by

$$\xi_t^{w,N}(A) = \{\text{total amount of jobs in system at time } t \text{ with their residual job size in } A\},$$

where $A \in \mathcal{B}(\mathbb{R})$. Also, define $\beta^{w,N} \in \mathbb{D}_{\mathcal{M}}^{\uparrow}$, as

$$\beta_t^{w,N}(A) = \{\text{total amount of work done by time } t \text{ on the jobs having initial job size in } A\},$$

for $A \in \mathcal{B}(\mathbb{R})$. We define $\gamma^N(x,t)$ to be the total amount of jobs present in the system at time t that have residual job size less than x and initial job size strictly bigger than x. Then one readily obtains the following balance equation

$$\xi_t^{w,N}[0,x] = \alpha_t^{w,N}[0,x] - \beta_t^{w,N}[0,x] - \gamma^N(t,x), \quad x \in \mathbb{R}_+,\ t \geq 0. \tag{5.72}$$

Note that for any $t, x > 0$, there could be at most one job present in the system at time t with residual job size less than x and initial job size strictly bigger than x. Thus,

$$\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\gamma}^N(t,x) \leq (\bar{\alpha}_T^{w,N} - \bar{\alpha}_{T-}^{w,N})(\mathbb{R}_+) \to 0, \quad \text{in probability}, \tag{5.73}$$

where the right-hand side follows from Assumption 5.10(1). On the other hand, from (5.56) and Step 1 above, we have

$$\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} |\bar{\xi}_t^{w,N}[0,x] - \xi_t^{*,N}[0,x]| \leq \|\bar{J}_2^{w,N}\|_T \to 0, \quad \text{in probability}. \tag{5.74}$$

On combining (5.58), (5.68), (5.72)-(5.74), we obtain

$$\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} |\bar{\beta}_t^{w,N}[0,x] - \bar{\beta}_t^{*,N}[0,x]| \to 0, \text{ in probability}.$$

Thus, by Step 4 above, we see that $(\bar{\xi}^{w,N}, \bar{\beta}^{w,N}) \Rightarrow \Theta(\alpha^w, \mu)$. It is also easily seen that one can analogously define $\xi^{n,N}, \beta^{n,N}$, associated to the job count process, and obtain a result similar to Theorem 5.16.

A Proof of Lemma 2.4

In this section, we give the proof of Lemma 2.4 which states various properties of $D_{\mathcal{M}}$. Let $\{\zeta^n\} \subset \mathbb{D}_{\mathcal{M}}$ be a sequence such that $\zeta^n \to \zeta$ for some $\zeta \in D_{\mathcal{M}}$. Then $\zeta_t^n \to \zeta_t$ in $\mathcal{M}$ at any point of continuity t of ζ. Thus, if $\zeta^n \in \mathbb{D}_{\mathcal{M}}^{\uparrow}$ for every $n \in \mathbb{N}$, and $t_1 < t_2$ are two continuity points of ζ, it follows by weak convergence that $0 \leq \langle f, \zeta_{t_1}\rangle \leq \langle f, \zeta_{t_2}\rangle$ for $f \in \mathbb{C}_{b,+}(\mathbb{R}_+)$. If t_1 (similarly, t_2) is not a continuity point, argue by selecting continuity points t^ℓ, such that $t^\ell \downarrow t_1$, and use the fact that $\zeta_{t^\ell} \to \zeta_{t_1}$ in $\mathcal{M}$ to deduce that $0 \leq \langle f, \zeta_{t_1}\rangle \leq \langle f, \zeta_{t_2}\rangle$. This shows that $\zeta \in \mathbb{D}_{\mathcal{M}}^{\uparrow}$, and hence that $\mathbb{D}_{\mathcal{M}}^{\uparrow}$ is a closed subset of $\mathbb{D}_{\mathcal{M}}(\mathbb{R}_+)$.

To establish property 2, fix $\zeta \in \mathbb{D}_{\mathcal{M}}^{\uparrow}$ and $0 \leq x < y$. Then by the definition of $\mathbb{D}_{\mathcal{M}}^{\uparrow}$, $\zeta[0,x]$ and $\zeta(x,y]$ are non-negative and non-decreasing in t. Let $t, t_n \in \mathbb{R}_+$ be such that $t_n \downarrow t$ as

$n \to \infty$. Since $\zeta \in \mathbb{D}_{\mathcal{M}}$, $\zeta_{t_n} \to \zeta_t$ as $n \to \infty$. Since $[0,x]$ is a closed subset of $\mathbb{R}_+$, by the Portmanteau theorem,

$$\limsup_{n\to\infty} \zeta_{t_n}[0,x] \le \zeta_t[0,x].$$

On the other hand, since $\zeta \in \mathbb{D}^{\uparrow}_{\mathcal{M}}$, by monotonicity, one has $\zeta_t[0,x] \le \zeta_{t_n}[0,x]$. As a result, $\lim_{n\to\infty} \zeta_{t_n}[0,x] = \zeta_t[0,x]$, showing that $\zeta[0,x]$ is a member of $\mathbb{D}_{\mathbb{R}}$. Since $\zeta(x,y) = \zeta[0,y] - \zeta[0,x]$, $\zeta(x,y]$ also lies in $\mathbb{D}_{\mathbb{R}}$. Combined with the monotonicity property proved earlier, this gives $\zeta[0,x] \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$ and $\zeta(x,y] \in \mathbb{D}^{\uparrow}_{\mathbb{R}}$. For the converse it is enough to show that $\langle f, \zeta_t \rangle$ is non-decreasing in t for every $f \in \mathbb{C}_{b,+}(\mathbb{R}_+)$ with compact support in $[0,\infty)$. Now any continuous function f with compact support can be approximated uniformly over $\mathbb{R}_+$ by functions of the form $f(0)\mathbb{I}_{[0,s_0]} + \sum_{i\ge1} f(s_i)\mathbb{I}_{(s_{i-1},s_i]}$ where $\{0 < s_0 < s_1 < \cdots\}$ forms a finite partition of $[0,\infty)$. Therefore if $\zeta[0,s_0]$ and $\zeta(s_{i-1},s_i]$ are non-decreasing, we have $\langle f, \zeta_t \rangle$ non-decreasing in t for $f \in \mathbb{C}_{b,+}(\mathbb{R}_+)$ with compact support.

We now turn to the proof of (2.4). Arguing by contradiction, assume there exist $\delta > 0$ and a sequence $\{s_n\} \subset [0,t]$ such that

$$\zeta_{s_n}(x,x_n] \ge \delta, \quad \text{for all } n. \tag{A.1}$$

Since the sequence $\{s_n\}$ lies in the compact set $[0,t]$, there exists $s \in [0,t]$ and a subsequence, which we denote again by $\{s_n\}$, such $s_n \to s$. By choosing a further subsequence, if necessary, we can assume that one of the following holds: either $s_n \uparrow s$ or $s_n \downarrow s$ as $n \to \infty$. If $s_n \uparrow s$ then using the monotonicity of $t \to \zeta_t$ in $\mathcal{M}$ and Lemma 2.4(2), we see that $0 \le \zeta_{s_n}(x,x_n] \le \zeta_s(x,x_n]$. Since $\zeta_s(x,x_n] \to 0$ as $n \to \infty$, this contradicts (A.1). Now, consider the case when $s_n \downarrow s$. Fix $\varepsilon > 0$. Then, for all sufficiently large n, we have $x_n < x + \varepsilon/2$ and, due to the right-continuity of $t \to \zeta_t$, we have $\zeta_{s_n}[0, x+\varepsilon/2] \le \zeta_{s_n}[0, x+\varepsilon) \le \zeta_s[0, x+\varepsilon) + \varepsilon/2$. Therefore, for all large n, using (A.1), the monotonicity of $t \mapsto \zeta_t$ and the above properties, we obtain

$$\begin{aligned} \delta \le \zeta_{s_n}[0,x_n] - \zeta_{s_n}[0,x] &\le \zeta_{s_n}[0,x_n] - \zeta_s[0,x] \\ &\le \zeta_{x_n}\left[0, x + \frac{\varepsilon}{2}\right] - \zeta_s[0,x] \\ &\le \zeta_s[0,x+\varepsilon] - \zeta_s[0,x] + \varepsilon/2 \\ &= \zeta_s(x, x+\varepsilon] + \varepsilon/2. \end{aligned}$$

Sending $\varepsilon \to 0$, the right-hand side goes to zero, which yields a contradiction. This proves the first limit in (2.4). The proof of the second limit is exactly analogous, and is thus omitted.

We turn to the proof of the last property. Since the Borel σ-field of $\mathbb{D}_{\mathcal{M}}$ is generated by finite dimensional projections, it suffices to show the measurability of the map $\mathcal{T}_t : (S,\mathcal{S}) \mapsto (\mathcal{M}, \mathcal{B}(\mathcal{M}))$, defined by $\mathcal{T}_t(s) = (\mathcal{T}(s))_t$. In turn, to show the latter, by the definition of the weak topology on $\mathcal{M}$, it suffices to show that for every $f \in \mathbb{C}_b(\mathbb{R}_+)$, the map $\mathcal{T}^f_t : (S,\mathcal{S}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by $\mathcal{T}^f_t(s) := \langle f, \mathcal{T}_t(s) \rangle$, is measurable for every $f \in \mathbb{C}_b(\mathbb{R}_+)$. Now, define $\mathfrak{H}_1 := \{\mathbb{I}_{[0,a]} \ : \ a \in \mathbb{R}_+\}$, $\bar{\mathfrak{H}}_1 := \{\mathbb{I}_{[0,\infty)}\} \cup \mathfrak{H}_1$ and

$$\mathfrak{H} := \{f \ : \ f \text{ is bounded, Borel measurable on } \mathbb{R}_+ \text{ and } T^f_t \text{ is also measurable}\}.$$

Thus, prove the lemma, it suffices to show that if $\mathfrak{H}_1 \subset \mathfrak{H}$, then $C_b(\mathbb{R}_+) \subset \mathfrak{H}$. If $\mathfrak{H}_1 \subset \mathfrak{H}$, then since $\mathbb{I}_{[0,\infty)} = \lim_{a\to\infty} \mathbb{I}_{[0,a]}$, the monotone convergence theorem shows that T^f_t is also measurable for $f = \mathbb{I}_{[0,\infty)}$, and hence, $\bar{\mathfrak{H}}_1 \subset \mathfrak{H}$. Clearly, $\mathfrak{H}$ is a vector space and hence, contains constants because $\bar{\mathfrak{H}}_1$ contains the function that is constant and equal to one, and $\bar{\mathfrak{H}}_1 \subset \mathfrak{H}$. Also, suppose f is bounded and $f_n \uparrow f$ pointwise for $f_n \in \mathfrak{H}$, $n \in \mathbb{N}$. Then the bounded convergence theorem shows that $T_t^f = \lim_{n\to\infty} T_t^{f_n}$ and hence, $f \in \mathfrak{H}$. Furthermore, $\bar{\mathfrak{H}}_1$ is closed under finite products. Hence, by the functional version of the monotone class theorem (see [10, Theorem 6.1.3]), $\mathfrak{H}$ contains all functions that are measurable with respect to the σ-field generated by $\bar{\mathfrak{H}}_1$. Since $\bar{\mathfrak{H}}_1$ generates the Borel σ-field on $\mathbb{R}_+$, $\mathfrak{H}$ contains all bounded Borel measurable functions on $\mathbb{R}_+$, and in particular, contains $\mathbb{C}_b(\mathbb{R}_+) \subset \mathfrak{H}$. This completes the proof of property 4. □

B Proof of Lemma 5.9

We shall work here with the filtration $\{\mathcal{F}_t\}$ obtained by augmenting in the usual way the filtration $\sigma\{\xi_s, \beta_s, \iota_s, \rho_s,\ s \le t\}$. The optional sets and processes defined below will be with respect to this filtration, and the measurable sets will be $\mathcal{G} := \mathcal{B}(\mathbb{R}_+) \times \mathcal{F}_\infty$-measurable where $\mathcal{B}(\mathbb{R}_+)$ are the Borel sets of $\mathbb{R}_+$.

We begin by showing that the set

$$\Gamma = \{(t,\omega) \in [0,T) \times \Omega : \sigma_t(\omega) > t + \delta,\ \rho(t+n^{-1}) > \rho(t) \text{ for all } n\}$$

is $\mathcal{G}$-measurable. For any $A \in \mathcal{B}[0,\infty)$, $\xi_t(\omega)(A)$ is $\{\mathcal{F}_t\}$-measurable. In particular for $A = [0, t+a]$, $\xi_t(\omega)[0,t+a]$ is $\{\mathcal{F}_t\}$-measurable. Further, as shown in Lemma 4.4,

$$t \mapsto \xi_t(\omega)[0,t+a] \text{ is right continuous.}$$

It follows that for each a, $(t,\omega) \mapsto \xi_t(\omega)[0,t+a]$ is optional and in particular, $\mathcal{G}$-measurable. As a result, the set

$$\{(t,\omega) \in [0,T) \times \Omega : \sigma_t(\omega) > t + \delta\} = \bigcup_{n=1}^{\infty} \{(t,\omega) : \xi_t[0, t+\delta+n^{-1}] = 0\}$$

is an optional set and therefore $\mathcal{G}$-measurable. Next, note that $\rho(t), t \ge 0$, is a continuous, adapted process, and thus $\mathcal{G}$-measurable. Hence $\rho(t+n^{-1}) - \rho(t)$ is $\mathcal{G}$-measurable for every n. It follows that Γ is $\mathcal{G}$-measurable.

By the Section Theorem for measurable sets (see e.g. Sharpe [33] p. 388 Theorem A5.8), there exists an $\mathcal{F}_\infty$-measurable random variable τ with values in $[0,T) \cup \{\infty\}$, so that $[\![\tau]\!] \subset \Gamma$, where $[\![\tau]\!]$ is the graph

$$\{(t,\omega) \in [0,T) \times \Omega : \tau(\omega) = t\},$$

and

$$\mathbb{P}(\tau < \infty) = \mathbb{P}(\text{there exists } t \text{ so that } (t,\omega) \in \Gamma).$$

Since the expression on the right-hand side is equal to $\mathbb{P}(E_0)$, the result follows. □

C Proof of Lemma 5.15

We now present the proof of Lemma 5.15. Fix $T < \infty$ and $\eta > 0$. For constants $k < \infty$ and $r_n, n \in \mathbb{N}$, chosen below, denote

$$\Omega_0^N := \{\zeta_T^N[0,\infty) < k\}, \qquad \Omega^{N,n} := \left\{\zeta_T^N(r_n,\infty) < \frac{1}{n}\right\}.$$

Recall that that a set $C \subset \mathcal{M}$ is relatively compact if $\sup_{\nu\in C}\nu(\mathbb{R}_+) < \infty$ and for every positive ε, there exists a compact set $K \subset \mathbb{R}_+$ such that $\sup_{\nu\in C}\nu(K^c) < \varepsilon$. Then, by the assumption that $\{\zeta_T^N\}$ satisfies (5.48), it follows that $k < \infty$ can be chosen so that, for every N, $\mathbb{P}(\Omega_0^N) > 1 - \eta/2$, and $r_n < \infty$, $n \in \mathbb{N}$, can be chosen so that $\mathbb{P}(\Omega^{N,n}) > 1 - 2^{-n-1}\eta$. Fix such k and $\{r_n\}$, and define $\Omega^N := \Omega_0^N \cap [\cap_{n\geq 1}\Omega^{N,n}]$. Then one has $\mathbb{P}(\Omega^N) > 1 - \eta$ for every N. Moreover, for every N, on the event Ω^N one has $\zeta_T^N \in K_{T,\eta}$, where

$$K_{T,\eta} := \{\nu \in \mathcal{M} : \nu(\mathbb{R}_+) < k \text{ and } \nu(r_n,\infty) < 1/n \text{ for all } n \in \mathbb{N}\}.$$

By (5.49) and the monotonicity of $t \mapsto \zeta_t^N$, we obtain

$$\mathbb{P}(\tilde{\zeta}_t^N \in K_{T,\eta} \text{ for all } t \in [0,T]) > 1 - \eta, \qquad \text{for all } N.$$

Note that

$$\inf_{\text{compact } C\subset\mathbb{R}_+} \sup_{\nu\in K_{T,\eta}} \nu(C^c) \leq \inf_n \sup_{\nu\in K_{T,\eta}} \nu((r(n),\infty)) = 0,$$

and

$$\sup_{\nu\in K_{T,\eta}} \nu[0,\infty) \leq k.$$

It follows that $K_{T,\eta}$ is relatively compact in $\mathcal{M}$, and we have thus shown that (5.48) holds for $\{\tilde{\zeta}^N\}$ with $\mathcal{K}_{T,\eta}$ equal to the closure of $K_{T,\eta}$ in the Levy metric. □

Acknowledgment. The research of RA was supported in part by grant 1315/12 of the Israel Science Foundation. The research of HK was supported in part by grant 764/13 of the Israel Science Foundation. The research of KR was supported in part by grant CMMI-1538706 of the National Science Foundation. The authors thank Subhamay Saha and an anonymous referee for comments on earlier versions of the manuscript.

References

[1] V. Anantharam and T. Konstantopoulos. Regulating functions on partially ordered sets. *Order*, 22:145–183, 2005.

[2] R. Atar, A. Biswas, and H. Kaspi. Fluid limits of $G/G/1+G$ queues under the non-preemptive earliest-deadline-first discipline. *Preprint*, 2013. URL `http://arxiv.org/abs/1206.5704`.

[3] R. Atar, H. Kaspi, and N. Shimkin. Fluid limits for many-server systems with reneging under a priority policy. *Math. Oper. Res.*, to appear, 2014.

[4] A. Biswas. Fluid limits of many-server queues with state dependent service rates. *Preprint*, 2014. URL `http://arxiv.org/abs/1206.5704`.

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

[6] H. Chen and A. Mandelbaum. Leontief systems, RBVs and RBMs. In *Applied stochastic analysis (London, 1989)*, volume 5 of *Stochastics Monogr.*, pages 1–43. Gordon and Breach, New York, 1991.

[7] L. Decreusefond and P. Moyal. Fluid limit of a heavily loaded EDF queue with impatient customers. *Markov Process. Related Fields*, 14(1):131–158, 2008. ISSN 1024-2953.

[8] D. G. Down, H. C. Gromoll, and A. L. Puha. Fluid limits for shortest remaining processing time queues. *Math. Oper. Res.*, 34(4):880–911, 2009. ISSN 0364-765X. URL `http://dx.doi.org/10.1287/moor.1090.0409`.

[9] B. Doytchinov, J. Lehoczky, and S. Shreve. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.*, 11(2):332–378, 2001. ISSN 1050-5164. URL `http://dx.doi.org/10.1214/aoap/1015345295`.

[10] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, fourth edition, 2010.

[11] S. N. Ethier and T. G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. ISBN 0-471-08186-8. x+534 pp. URL `http://dx.doi.org/10.1002/9780470316658`. Characterization and convergence.

[12] C. Gromoll. Diffusion approximation of a processor sharing queue in heavy traffic. *Ann. Appl. Probab.*, 14 (2):555–611, 2004.

[13] C. Gromoll, L. Kruk, and A. Puha. Diffusion limits for shortest remaining processing time queues. *Stoch. Systems*, 1(1):1–16, 2011.

[14] H. C. Gromoll and M. Keutel. Invariance of fluid limits for the shortest remaining processing time and shortest job first policies. *Queueing Syst.*, 70(2):145–164, 2012. ISSN 0257-0130. URL `http://dx.doi.org/10.1007/s11134-011-9267-5`.

[15] H. C. Gromoll, A. L. Puha, and R. J. Williams. The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.*, 12(3):797–859, 2002. ISSN 1050-5164. URL `http://dx.doi.org/10.1214/aoap/1031863171`.

[16] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009. ISBN 978-0-470-12990-6. xvi+354 pp. + loose erratum pp. URL `http://dx.doi.org/10.1002/9780470434697`.

[17] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003. ISBN 3-540-43932-3. xx+661 pp. URL `http://dx.doi.org/10.1007/978-3-662-05265-5`.

[18] A. Jakubowski. On the Skorokhod topology. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(3):263–285, 1986. CODEN AHPBAR. ISSN 0246-0203. URL `http://www.numdam.org/item?id=AIHPB_1986__22_3_263_0`.

[19] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.*, 20(6): 2204–2260, 2010. ISSN 1050-5164. URL `http://dx.doi.org/10.1214/10-AAP683`.

[20] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Ann. Appl. Probab.*, 21 (1):33–114, 2011. ISSN 1050-5164. URL `http://dx.doi.org/10.1214/09-AAP662`.

[21] H. Kaspi and K. Ramanan. SPDE limits of many-server queues. *Ann. Appl. Probab.*, 23(1):145–229, 2013. ISSN 1050-5164. URL `http://dx.doi.org/10.1214/11-AAP821`.

[22] Ł. Kruk. Stability of two families of real-time queueing networks. *Probab. Math. Statist*, 28:179–202, 2008.

[23] Ł. Kruk et al. Invariant states for fluid models of edf networks: Nonlinear lifting map. *Probab. Math. Statist., v30*, pages 289–315, 2010.

[24] L. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. An explicit formula for the Skorokhod map on $[0, a]$. *Ann. Probab.*, 35(5):1740–1768, 2007.

[25] L. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. *Double Skorokhod map and reneging real-time queues*, volume 4 of *IMS Collections*, pages 169–193. IMS, 2008. Preprint, to appear in a Festschrift Volume for Tom Kurtz.

[26] Ł. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Heavy traffic analysis for EDF queues with reneging. *Ann. Appl. Probab.*, 21(2):484–545, 2011. ISSN 1050-5164. URL `http://dx.doi.org/10.1214/10-AAP681`.

[27] P. Moyal. On queues with impatience: stability, and the optimality of earliest deadline first. *Queueing Syst.*, 75(2-4):211–242, 2013. ISSN 0257-0130. URL `http://dx.doi.org/10.1007/s11134-013-9342-1`.

[28] S. S. Panwar and D. Towsley. On the optimality of the ste rule for multiple server queues that serve customer with deadlines. *Technical Report 88–81*, pages Dept. of Computer and Information Science, Univ. Massachusetts, Amherst., 1988.

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

[30] A. Puha. Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. *Ann. Appl. Probab.*, 25(6):3381–3404, 2015.

[31] K. Ramanan. Reflected diffusions defined via the extended Skorokhod map. *Elec. Jour. Probab.*, 11: 934–992, 2006.

[32] A. Sharif, D. Stanford, and I. Ziedins. A multi-class multi-server accumulating priority queue with application to health care. *Operations Research for Health Care*, 3(2):73–79, 2014.

[33] M. Sharpe. *General theory of Markov processes*, volume 133 of *Pure and Applied Mathematics*. Academic Press, Inc., Boston, MA, 1988. ISBN 0-12-639060-6. xii+419 pp.

[34] A. Skorokhod. Stochastic equations for diffusions in a bounded region. *Theor. of Prob. and Appl.*, 6: 264–274, 1961.

[35] D. Stanford, P. Taylor, and I. Ziedins. Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77(3):297–330, 2014.

[36] J. Zhang, J. G. Dai, and B. Zwart. Law of large number limits of limited processor-sharing queues. *Math. Oper. Res.*, 34(4):937–970, 2009. ISSN 0364-765X. URL `http://dx.doi.org/10.1287/moor.1090.0412`.