# A fluid approximation for a matching model with general reneging distributions

Angelos Aveklouris[1], Amber L. Puha[2], and Amy R. Ward[1]

[1]The University of Chicago Booth School of Business
[2]California State University San Marcos

August 31, 2023

**Abstract**

Motivated by a service platform, we study a two-sided network where heterogeneous demand (customers) and heterogeneous supply (workers) arrive randomly over time to get matched. Customers and workers arrive with a randomly sampled patience time (also known as reneging time in the literature), and are lost if forced to wait longer than that time to be matched. The system dynamics depend on the matching policy, which determines when to match a particular customer class with a particular worker class. Matches between classes use the head-of-line customer and worker from each class. Since customer and worker arrival processes can be very general counting processes, and the reneging times can be sampled from any finite mean distribution that is absolutely continuous, the state descriptor must track the age-in-system for every customer and worker waiting in order to be Markovian, as well as the time elapsed since the last arrival for every class. We develop a measure-valued fluid model that approximates the evolution of the discrete-event stochastic matching model, and prove its solution is unique under a fixed matching policy. For a sequence of matching models, we establish a tightness result for the associated sequence of fluid-scaled state descriptors, and show that any distributional limit point is a fluid model solution almost surely. When arrival rates are constant, we characterize the invariant states of the fluid model solution, and show convergence to these invariant states as time becomes large. Finally, again when arrival rates are constant, we establish another tightness result for the sequence of fluid-scaled state descriptors distributed according to a stationary distribution, and show that any subsequence converges to an invariant state. As a consequence, the fluid and time limits can be interchanged, which justifies regarding invariant states as first order approximations to stationary distributions.

**Keywords**: service platforms; two-sided platform; reneging; fluid approximation; functional limit theorems; measure-valued process

## 1 Introduction

Service platforms appear in many applications (e.g., ridesharing, online marketplaces, etc.) and have to match demand (customers) and supply (workers) taking into account their heterogeneity, the random arrival times, and the random impatience of customers and workers. The objective of a platform is to consider a matching policy (i.e., when to make matches and between whom) to

optimize the performance of the system. For example, the platform may want to maximize the cumulative value of matches made, minimize the loss of customers and workers, minimize possible holding costs or a combination of the aforementioned objectives. For this, a platform needs to know the demand and supply waiting to be matched (i.e., queue-lengths), and how that evolves over time.

We model a service platform as a two-sided graph where an arbitrary number of customer and worker types arrive randomly to each side in order to be matched, according to arrival rates that may vary over time. Each customer and worker arrives with a patience time randomly sampled from a type-dependent distribution with finite mean, and is lost if not matched within the patience time. A Markovian state descriptor must track the age-in-system for every customer and worker waiting, and so is measure-valued. As a result, exact analysis of our model appears intractable.

Our focus in this work is to provide a fluid approximation for this system when matchings occur between head-of-the-line (HL) customers and workers, to characterize the fluid invariant states when arrival rates are constant, and to establish rigorous convergence results to support the fluid approximation. The analytic tractability of the fluid approximation provides a framework that the platform can use to choose the "correct" matching policy that optimizes the performance of the system (but which is studied in the companion paper [6]).

The main contributions of this paper are:

(1) *Non-policy-specific fluid limits.* We provide a tightness result for a sequence of matching models, that holds without the need to fully specify the matching policy (see Theorem 2). Then, we prove that any subsequential limit is almost surely a fluid model solution (see Theorem 3).

(2) *Uniqueness of fluid limits.* We establish that a fluid limit is unique under a fixed matching policy (see Theorem 1).

(3) *Convergence of stationary distributions.* When arrival rates are constant, we show a tightness result for a sequence of matching models operating in stationarity, and prove that any subsequential limit is a fluid model invariant state (see Theorem 4).

(4) *Interchange of Limits.* Theorems 2 and 4, combined with results on stationary distribution existence (Proposition 1), on characterization of fluid invariant states (Proposition 2), and on convergence to fluid invariant states (Proposition 3), imply an interchange of limits (illustrated in Figure 2) that justifies regarding invariant states as first order approximations to stationary distributions.

Our proofs heavily leverage the methodology developed in [29] and [30] for a single-class many-server queue with reneging, and in [5] and [43] for a multiclass many-server queue with reneging. All four aforementioned papers make clever use of a what is termed a "potential queue measure", that stores the amount of time that has passed since each customer's arrival time, up until the customer's patience time. The potential queue measure greatly facilitates analysis because the measure does not depend on the policy for serving customers. Similarly, we use a potential queue measure that does not depend on the policy for matching customers and workers, which allows us to leverage many results in [29], [30] and [43] to prove item (1) above. Differently, because matching is instantaneous in our model (so that there is no equivalent of service time in the queueing framework), the long-time behavior of our fluid model is easier to analyze than the fluid model relevant to the multiclass $G/GI/N+GI$ queue in [5] and [43], which is key to some of the proofs for items (3) and (4) above.

Moreover, to prove item (2) above that a fluid model solution is unique under a specified matching policy, we do not need to assume that the hazard rates associated with the reneging distributions are bounded, as is needed in the scheduling policies for the multiclass $G/GI/N+GI$ queues analyzed in [5] and [43].

### Some Related Literature

Queueing systems with primitive inputs that follow distributions that are not exponential have complicated state descriptors that motivate the use of measure-valued processes. In addition to the papers mentioned in the previous paragraph, some other examples of papers that use measure-valued state descriptors to study many server queueing systems with reneging customers are [5, 47, 51]. Other queueing situations in which measure-valued state descriptors are used include LIFO queues [35], SRPT queues [7, 19, 23, 42], many-server retrial queues with nonpersistent customers [28], processor-sharing queues in [44, 48], processor-sharing queues with impatient customers [21], load-balancing algorithms [2, 3], and bandwidth-sharing networks [22, 45].

Our work is related to work that studies service platforms. These platforms fit into the sharing economy; see [9, 16, 24, 25] for perspectives and research opportunities in this area through the lens of operations management. From that perspective, our two-sided matching model with reneging can be viewed as a model of a service platform. There are many works on two-sided matching models, less on two-sided matching models with reneging. The works [1, 4, 8, 12, 13, 14, 15, 18, 27, 31, 33, 34, 36, 39, 40, 41, 46, 50] include reneging, but all assume that reneging times are either deterministic or exponentially distributed. Like us, [13, 18, 33] allow for more general reneging distributions; however, [13, 33] restricts to one demand and one supply type, and [18] focuses on one specific policy class (an index policy class).

### Organization of the Paper

The remainder of the paper is organized as follows. We end this section by summarizing our mathematical notation. Section 2 specifies our detailed discrete-event stochastic matching model. We provide the fluid model equations in Section 3, and establish a uniqueness result (Theorem 1). We provide a non-policy specific tightness result (Theorem 2) and a convergence result (Theorem 3) in Section 4. Finally, in Section 5, under the assumption that arrival rates are constant, we provide a convergence result for a sequence of matching models operating in stationarity (Theorem 4).

### Notation

We use the following notational conventions. All vectors and matrices are denoted by bold letters. Further, $\mathbb{R}$ is the set of real numbers, $\mathbb{R}_+$ is the set of nonnegative real numbers, $\mathbb{N}$ is the set of strictly positive integers, and $\mathbb{Z}_+ = \mathbb{N} \cup \{0\}$. The sets $\mathbb{R}$ and $\mathbb{R}_+$ are endowed with the Euclidean topology, and $\mathbb{Z}_+$ with the discrete topology. Moreover, for $m \in \mathbb{N}$ and a vector $\boldsymbol{x} \in \mathbb{R}^m$, $\|\boldsymbol{x}\|_\infty := \max_{1 \leqslant i \leqslant m} |x_i|$ is the maximum norm.

For a measurable space $(S, \mathcal{F})$ and a measurable set $A \in \mathcal{F}$, $1_A$ is the indicator function of the set $A$, which is one when its argument is a member of the set $A$ and is zero otherwise. In addition, when $A$ is $S$, we use the shorthand notation 1 to mean $1_S$. Also, for $\mathcal{A} \subset \mathcal{F}$, $\sigma(\mathcal{A})$ denotes the $\sigma$-algebra generated by $\mathcal{A}$.

Let $H \in (0,\infty]$. Then, $\mathcal{C}_c([0,H))$ (resp. $\mathcal{C}_b([0,H))$) denotes the set of continuous, compactly supported (resp. bounded) functions $f : [0,H) \to \mathbb{R}$, and $\mathcal{C}_c^1([0,H))$ denotes the set of continuous, compactly supported functions $f : [0,H) \to \mathbb{R}$ for which the derivative $f'$ exists for all $x \in [0,H)$ and $t \geqslant 0$, and lies in $\mathcal{C}_c([0,H))$. Similarly, $\mathcal{C}_c([0,H) \times \mathbb{R}_+)$ (resp. $\mathcal{C}_b([0,H) \times \mathbb{R}_+)$) denotes the set of continuous, compactly supported (resp. bounded) functions $\varphi : [0,H) \times \mathbb{R}_+ \to \mathbb{R}$, and $\mathcal{C}_c^{1,1}([0,H) \times \mathbb{R}_+)$ denotes the set of continuous, compactly supported functions $\varphi : [0,H) \times \mathbb{R}_+ \to \mathbb{R}$ for which the directional derivative $\lim_{\varepsilon \to 0} \frac{\varphi(x+\varepsilon,t+\varepsilon)-\varphi(x,t)}{\varepsilon}$ exists for all $x \in [0,H)$ and $t \geqslant 0$, and lies in $\mathcal{C}_c([0,H),\mathbb{R}_+)$. We shall abuse the notation by using $\varphi_x + \varphi_t$ to denote this directional derivative, whether the partial derivatives exist or not. Finally, $\mathbf{L}^1([0,H))$ (resp. $\mathbf{L}^1_{\mathrm{loc}}([0,H))$) denotes the set of Borel measurable functions on $[0,H)$ that are integrable (resp. locally integrable) with respect to Lebesgue measure on $[0,H)$.

Given a Polish space $\mathbb{S}$, we use the notation $\mathcal{C}(\mathbb{S})$ (with no subscript) to denote the set of $\mathbb{S}$ valued functions with domain $\mathbb{R}_+$ that are continuous, and the notation $\mathcal{D}(\mathbb{S})$ to denote the set of $\mathbb{S}$ valued functions with domain $\mathbb{R}_+$ that are right continuous with finite left limits (rcll). We endow $\mathcal{C}(\mathbb{S})$ and $\mathcal{D}(\mathbb{S})$ with the usual Skorokhod $J_1$-topology [10]. In contrast to the sets of functions defined in the previous paragraph, we use the range rather than the domain as the argument. The domain is always time, $\mathbb{R}_+$.

For $L \in [0,\infty]$, let $\mathcal{M}[0,L)$ denote the set of finite, non-negative Borel measures on $[0,L)$ endowed with the topology of weak convergence, which is a Polish space. Given a measure $\nu \in \mathcal{M}[0,L)$ and a Borel measurable function $f : [0,L] \to \mathbb{R}$ that is integrable with respect to $\nu$ define $\langle f, \nu \rangle := \int_{[0,L)} f(x)\nu(dx)$. Given $x \in [0,H)$, $\delta_x \in \mathcal{M}[0,L)$ is the Dirac measure with unit atom at $x$, i.e., for all Borel measurable $A \subset [0,H), \langle 1_A, \delta_x \rangle = 1_A(x)$.

Given a cumulative distribution function $G$ defined on $\mathbb{R}_+$ that is absolutely continuous with respect to Lebesgue measure and having probability density function $g$, the right edge of its support is given by

$$H = \sup\{x \in \mathbb{R}_+ : G(x) < 1\} \in (0,\infty]$$

Let $h$ denote the associated hazard function; i.e., $h(x) = \frac{g(x)}{1-G(x)}$ with $x \in [0,H)$. Then, $h \in \mathbf{L}^1_{\mathrm{loc}}([0,H))$. To see this, note that by assumption $G$ is absolutely continuous on $\mathbb{R}_+$ and, since ln is Lipschitz continuous on $[a,\infty)$ for any $a > 0$, it follows that $-\ln(1-G(x))$, $x \in [0,H)$, is absolutely continuous on $[0,b]$ for any $b < H$.

Given a counting process $A$, i.e., a nondecreasing integer valued process such that $A(0) = 0$, $A(t) < \infty$ for all $t \geqslant 0$ and $\lim_{t\to\infty} A(t) = \infty$, the jump times $(e_i)_{i\in\mathbb{N}}$ are given by

$$e_i = \inf\{t \geqslant 0 : A(t) \geqslant i\}, \qquad i \in \mathbb{N}.$$

Then $(e_i)_{i\in\mathbb{N}}$ is a nondecreasing sequence such that $\lim_{i\to\infty} e_i = \infty$. If the counting process has jumps of size one, $(e_i)_{i\in\mathbb{N}}$ is strictly increasing. The associated age process $a$, also known as the backward recurrence time process, is such that given $\alpha \in \mathbb{R}_+$,

$$a(t) = \begin{cases} \alpha + t, & t \in [0, e_1), \\ t - \sup\{s < t : A(t) - A(s) > 0\}, & t \geqslant e_1. \end{cases}$$

Then, for each $t \geqslant 0$, $a(t)$ represents the age of (the time that has elapsed since) the most recent jump event to occur at or before time $t$ happened. When the jumps are of size one, the age process uniquely determines the counting process. Otherwise, some information about the jump sizes is required. With a slight abuse of language, we will say that a counting process $A$ is a Markov counting process if the process $(a, A)$ is a Markov process with respect to its own natural filtration.

## 2 Model description

There is a set of $J$ demand buffers $\mathbb{J} := \{1, \ldots, J\}$ (representing customer types) and a set of $K$ supply buffers $\mathbb{K} := \{1, \ldots, K\}$ (representing worker types) as shown in Figure 1. Customers and workers arrive randomly over time to the system, either individually or in batches, and are placed in the buffer for their type to be matched. There, they wait in first come, first served (FCFS) order, so that the head-of-the-line (HL) customer or worker of a given type is the one that has been waiting the longest. Customers and workers arrive with a patience time of random length, and are lost if not matched within their patience time. A matching policy specifies when to match customers and workers of different types, and always matches the HL customer and worker within each type. The set $\mathcal{E} \subseteq \mathbb{J} \times \mathbb{K}$ denotes the set of compatible matches between demand and supply nodes, i.e., demand type $j \in \mathbb{J}$ can be matched with supply type $k \in \mathbb{K}$ if and only if $(j, k) \in \mathcal{E}$.

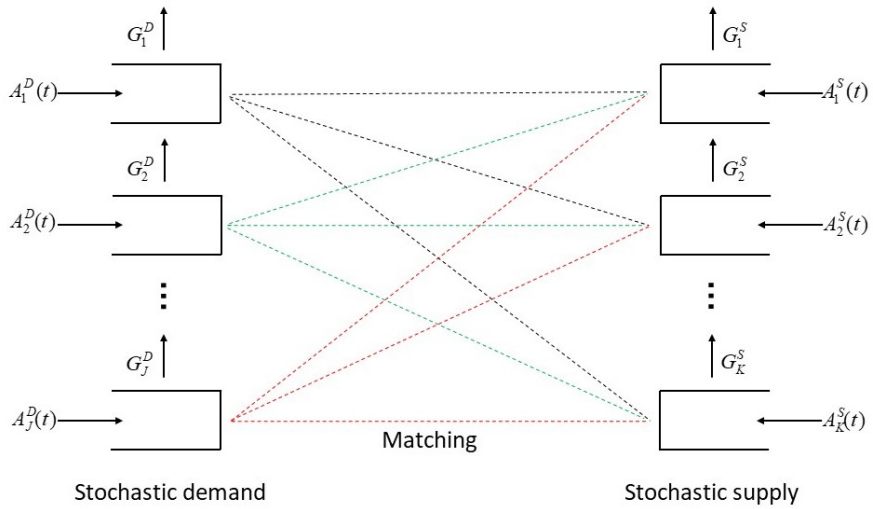


Figure 1: The two-sided matching model with general reneging distributions.

In what follows, we give a detailed model description. Throughout, we regard all random elements as being defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation operator $\mathbb{E}$. Section 2.1 provides the model inputs. Section 2.2 specifies the state descriptor and the system dynamics. Section 2.3 defines an admissible matching policy.

### 2.1 The Model Inputs

The model inputs consist of the arrival processes and stochastic primitives, which we define here.

#### 2.1.1 The Arrival Processes

We assume that demand of type $j \in \mathbb{J}$ and supply of type $k \in \mathbb{K}$ arrive according to Markov counting processes, denoted by $A_j^D$ and $A_k^S$ with age processes denoted by $a_j^D$ and $a_k^S$ respectively. The arrival time of the $l$th type $j$ customer and the arrival time of the $h$th type $k$ worker can respectively be expressed as

$$e_{jl}^D = \inf\{t \geqslant 0 : A_j^D(t) \geqslant l\}, j \in \mathbb{J}, l \in \mathbb{N} \qquad \text{and} \qquad e_{kh}^S = \inf\{t \geqslant 0 : A_k^S(t) \geqslant h\}, k \in \mathbb{K}, h \in \mathbb{N}.$$

Customers and workers may arrive one at a time, in which case the jump sizes of the arrival processes are one, or they may arrive in batches, in which case the jump sizes are positive integers. The arrival processes $\boldsymbol{A}^D$ and $\boldsymbol{A}^S$ are assume to be mutually independent of one another, and so are the coordinate processes. We further assume that for all $\boldsymbol{\alpha}^D \in (0,\infty)^J$, $\boldsymbol{\alpha}^S \in (0,\infty)^K$ and $t \geqslant 0$,

$$\max_{j\in\mathbb{J}} \mathbb{E}\left[A_j^D(t) \mid a_j^D(0) = \alpha_j^D\right] < \infty \qquad \text{and} \qquad \max_{k\in\mathbb{K}} \mathbb{E}\left[A_k^S(t) \mid a_k^S(0) = \alpha_k^S\right] < \infty. \tag{1}$$

### 2.1.2 The Stochastic Primitives

We denote the patience time of the $l$th type $j$ customer and the patience time of the $h$th type $k$ worker by $r_{jl}^D$ and $r_{kh}^S$, respectively. If an arriving customer or worker is not matched within their patience time, then that customer or worker reneges (abandons the system without being matched). Upon arrival, each type $j \in \mathbb{J}$ customer independently samples from the distribution determined by cumulative distribution function (cdf) $G_j^D$ to determine his patience time. Similarly, upon arrival, each type $k \in \mathbb{K}$ worker independently samples from the distribution determined by cdf $G_k^S$ to determine his patience time. We refer to $G_j^D, j \in \mathbb{J}$ and $G_k^S, k \in \mathbb{K}$ as the *reneging distributions* (also known as *patience time distributions*). Further, for each $j \in \mathbb{J}$, $G_{jy}^D$ (resp. for each $k \in \mathbb{K}$ $G_{ky}^S$) denotes the conditional cdf associated with $G_j^D$ (resp. $G_k^S$) conditioned to exceed $y \in [0, H_j^D)$ (resp. $y \in [0, H_k^S)$), where $H_j^D \in [0,\infty]$ (resp. $H_k^S \in [0,\infty]$) is the right edge of the support of the customer class $j$ reneging distribution (resp. supply class $k$). We assume the patience times are absolutely continuous random variables with density functions $g_j^D$ for $j \in \mathbb{J}$ and $g_k^S$ for $k \in \mathbb{K}$ that are mutually independent of each other, and of the arrival processes $\boldsymbol{A}^D$ and $\boldsymbol{A}^S$.

Finally, for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, let $\left\{U_{jl}^D\right\}_{l\in\mathbb{N}}$ and $\left\{U_{kh}^S\right\}_{h\in\mathbb{N}}$ be i.i.d sequences of uniform $(0,1)$ random variables that are mutually independent of one another, the arrival processes $\boldsymbol{A}^D$ and $\boldsymbol{A}^S$, and the patience times $\left\{r_{jl}^D\right\}_{l\in\mathbb{N}}, j \in \mathbb{J}$, and $\left\{r_{kh}^S\right\}_{h\in\mathbb{N}}, k \in \mathbb{K}$. These will be used to define various residual times associated with the initial condition.

We refer to the collection of sequences $\left\{U_{jl}^D\right\}_{l\in\mathbb{N}},\ j \in \mathbb{J}$, $\left\{U_{kh}^S\right\}_{h\in\mathbb{N}},\ k \in \mathbb{K}$, $\left\{r_{jl}^D\right\}_{l\in\mathbb{N}},\ j \in \mathbb{J}$, and $\left\{r_{kh}^S\right\}_{h\in\mathbb{N}},\ k \in \mathbb{K}$, as the stochastic primitives.

## 2.2 State descriptor and system dynamics

In the following, we first discuss the state space, then discuss the system dynamics that are independent of the matching decisions, and, finally, provide the evolution equations for the system processes that depend on the matching decisions.

### 2.2.1 System state

A state in our model is a vector $\boldsymbol{y} := \left(\boldsymbol{\alpha}^D, \boldsymbol{\alpha}^S, \boldsymbol{q}^D, \boldsymbol{q}^S, \boldsymbol{\eta}^D, \boldsymbol{\eta}^S\right) \in \mathbb{Y}_0$ where

$$\mathbb{Y}_0 := \mathbb{R}_+^J \times \mathbb{R}_+^K \times \mathbb{Z}_+^J \times \mathbb{Z}_+^K \times \left(\times_{j=1}^J \mathcal{M}[0, H_j^D)\right) \times \left(\times_{k=1}^K \mathcal{M}[0, H_k^S)\right).$$

It is known that the set $\mathbb{Y}_0$ endowed with the topology of weak convergence is a Polish space; see [11]. We now give an informal explanation of the state descriptor. Suppose the system is in state $\boldsymbol{y} \in \mathbb{Y}_0$. For $j \in \mathbb{J}$ and $k \in \mathbb{K}$, the quantities $\alpha_j^D$ and $\alpha_k^S$ denote the time that has elapsed since the last type $j$ batch of customers and last type $k$ batch of workers arrived to the system. Further,

for $j \in \mathbb{J}$ and $k \in \mathbb{K}$, $q_j^D$ and $q_k^S$ denote the number of customers and workers at queues $j$ and $k$, respectively. For each $j \in \mathbb{J}$, the measure $\eta_j^D \in \mathcal{M}[0, H_j^D)$ stores the amount of time that has passed between each type $j$ customer's arrival time up until that customer's potential abandonment time (the arrival time plus the sampled patience time). More specifically, for every type $j$ customer that has arrived by a given time, and whose potential abandonment time is after that time, $\eta_j^D$ has a unit atom supported at the potential waiting time of that class $j$ customer (the time that has elapsed since that class $j$ customer arrived). For each $k \in \mathbb{K}$, the measure $\eta_k^S \in \mathcal{M}[0, H_k^S)$ has analogous interpretation. This is without regard for whether that customer or worker has been matched, meaning these are the potential customers in queue. The FCFS matching assumption implies that all potential customers that have waited longer than the customer at the head-of-the-line have already been matched, and all potential customers that have waited less than that customer are in queue.

The system state will be an element of $\mathbb{Y}_0$ for all time, and will additionally be such that the number of customers in queue never exceeds the number of customers potentially in queue. Specifically, we are interested in the subset of $\mathbb{Y}$ of $\mathbb{Y}_0$ consisting of all $\boldsymbol{y} \in \mathbb{Y}_0$ such that

$$q_j^D \leqslant \langle 1, \eta_j^D \rangle \text{ and } q_k^S \leqslant \langle 1, \eta_k^S \rangle \text{ for each } j \in \mathbb{J}, k \in \mathbb{K}. \tag{2}$$

Note that in a slight abuse of notation we use $\boldsymbol{\eta}^D$ and $\boldsymbol{\eta}^S$ to represent a component of the system state but in what follows we use $\boldsymbol{\eta}^D$ and $\boldsymbol{\eta}^S$ to represent a process whose value at time $t \geqslant 0$ is a component of the system state.

### 2.2.2 Potential queue measures and potential reneging processes

Here we define the potential queue measures more formally. We begin with their initial value $(\boldsymbol{\eta}^D(0), \boldsymbol{\eta}^S(0)) \in \times_{j=1}^J \mathcal{M}[0, H_j^D) \times \times_{k=1}^K \mathcal{M}[0, H_k^S)$. For each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, there are $\langle 1, \eta_j^D(0) \rangle \in \mathbb{Z}_+$ type $j$ potential customers and $\langle 1, \eta_k^S(0) \rangle \in \mathbb{Z}_+$ type $k$ potential workers that arrived at or prior to time zero whose potential abandonment time is after time zero. Let $0 \leqslant w_{jl}^D(0) < H_j^D$ (resp. $0 \leqslant w_{kh}^S(0) < H_k^S$) for $l = -\langle 1, \eta_j^D(0) \rangle + 1, \ldots, 0$ (resp. $h = -\langle 1, \eta_k^S(0) \rangle + 1, \ldots, 0$) be the amount of time that has elapsed since type $j$ potential initial customer $l$ (type $k$ potential initial worker $h$) arrived. For each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, we assume that the sequences $\{w_{jl}^D(0)\}_{l=-\langle 1, \eta_j^D(0) \rangle + 1}^0$ and $\{w_{kh}^S(0)\}_{-\langle 1, \eta_k^S(0) \rangle + 1}^0$ are non-increasing in $l$ and $h$, respectively, and set $e_{jl}^D = -w_{jl}^D(0)$ for $l = -\langle 1, \eta_j^D(0) \rangle + 1, \ldots, 0$ and $e_{kh}^S = -w_{kh}^S(0)$ for $h = -\langle 1, \eta_k^S(0) \rangle + 1, \ldots, 0$. Then, for $j \in \mathbb{J}$ and $k \in \mathbb{K}$,

$$\eta_j^D(0) = \sum_{l=-\langle 1, \eta_j^D(0) \rangle + 1}^{0} \delta_{w_{jl}^D(0)} \text{ and } \eta_k^S(0) = \sum_{h=-\langle 1, \eta_k^S(0) \rangle + 1}^{0} \delta_{w_{jl}^D(0)}.$$

Next we define the patience times for the customers and workers in system at time 0. For $l = -\langle 1, \eta_j^D(0) \rangle + 1, \ldots, 0$ and $h = -\langle 1, \eta_k^S(0) \rangle + 1, \ldots, 0$, noting that any customer or worker present at time 0 must have patience time exceeding the amount of time that has passed since his or her arrival, the patience times of type $j$ zero potential customer $l$ and type $k$ zero potential worker $h$ are given by

$$r_{jl}^D = \inf \left\{ t > 0 : G_{jw_{jl}^D(0)}^D(t) > U_{jl}^D \right\} + w_{jl}^D(0)$$

and

$$r^S_{kh} = \inf\left\{t > 0 : G^S_{kw^S_{kh}(0)}(t) > U^S_{kh}\right\} + w^S_{kh}(0),$$

recalling that $U^D_{jl}$ and $U^S_{kh}$ are uniform $(0,1)$ random variables.

Finally, we define $(\boldsymbol{\eta}^D(t), \boldsymbol{\eta}^S(t))$ for $t > 0$. For this, we must define the potential waiting times. For each $l = \{-\langle 1, \eta^D_j(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$ and $t \geqslant 0$, the potential waiting time of the $l$th type $j$ potential customer at time $t \geqslant 0$ is given by

$$w^D_{jl}(t) := \min\left\{[t - e^D_{jl}]^+, r^D_{jl}\right\}.$$

In an analogous way for each $h = \{-\langle 1, \eta^S_k(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$, we define the potential waiting time of the $h$th type $k$ potential worker at time $t \geqslant 0$ as

$$w^S_{kh}(t) := \min\left\{[t - e^S_{kh}]^+, r^S_{kh}\right\}.$$

For any $t \geqslant 0$, $j \in \mathbb{J}$ and any Borel measurable $B \subseteq [0, H^D_j)$, let

$$\eta^D_j(t)(B) := \sum_{l=-\left\langle 1,\eta^D_j(0)\right\rangle+1}^{A^D_j(t)} \delta_{w^D_{jl}(t)}(B) 1_{\{0 \leqslant t - e^D_{jl} < r^D_{jl}\}}, \tag{3}$$

and, for any $t \geqslant 0$, $k \in \mathbb{K}$ and any Borel measurable $B \subseteq [0, H^S_k)$,

$$\eta^S_k(t)(B) := \sum_{h=-\left\langle 1,\eta^S_k(0)\right\rangle+1}^{A^S_k(t)} \delta_{w^S_{kh}(t)}(B) 1_{\{0 \leqslant t - e^S_{kh} < r^S_{kh}\}}. \tag{4}$$

Then, for each $t \geqslant 0$ and $j \in \mathbb{J}$, $\langle 1, \eta^D_j(t)\rangle$ is the number of type $j$ potential customers in the queue that arrived by time $t$ and whose potential waiting time is less than their patience time. Note that at time $t$ such customers may be in queue waiting to be matched or may have been matched and departed the system. For each $t \geqslant 0$ and $k \in \mathbb{K}$, $\langle 1, \eta^S_k(t)\rangle$ has an analogous meaning. By definition, for all $t \geqslant 0$,

$$\langle 1, \eta^D_j(t)\rangle \leqslant \langle 1, \eta^D_j(0)\rangle + A^D_j(t) \text{ for } j \in \mathbb{J}, \tag{5}$$

and

$$\langle 1, \eta^S_k(t)\rangle \leqslant \langle 1, \eta^S_k(0)\rangle + A^S_k(t) \text{ for } k \in \mathbb{K}. \tag{6}$$

Collections of marked point processes are used to characterize the dynamic evolution of the potential queue processes $\boldsymbol{\eta}^D$ and $\boldsymbol{\eta}^S$. To this end, for each $j \in \mathbb{J}$, $k \in \mathbb{K}$, measurable function $\phi : [0, H^D_j) \times \mathbb{R}_+ \to \mathbb{R}_+$, and measurable function $\psi : [0, H^S_k) \times \mathbb{R}_+ \to \mathbb{R}_+$, define the marked point processes $\mathcal{S}^D_j(\phi, \cdot)$ and $\mathcal{S}^S_k(\psi, \cdot)$ for $t \geqslant 0$,

$$\mathcal{S}^D_j(\phi, t) := \sum_{l=-\left\langle 1,\eta^D_j(0)\right\rangle+1}^{A^D_j(t)} \sum_{s\in[0,t]} 1_{\{\frac{dw^D_{jl}}{dt}(s-)>0, \frac{dw^D_{jl}}{dt}(s+)=0\}} \phi(w^D_{jl}(s), s), \tag{7}$$

$$\mathcal{S}^S_k(\psi, t) := \sum_{h=-\left\langle 1,\eta^S_k(0)\right\rangle+1}^{A^S_k(t)} \sum_{s\in[0,t]} 1_{\{\frac{dw^S_{kh}}{dt}(s-)>0, \frac{dw^S_{kh}}{dt}(s+)=0\}} \psi(w^S_{kh}(s), s). \tag{8}$$

When the functions $\phi$ and $\psi$ are replaced by the indicator of $\mathbb{R}_+$, we get the potential cumulative reneging processes; i.e., for $t \geqslant 0$, $S_j^D(t) := \mathcal{S}_j^D(1,t), j \in \mathbb{J}$ and $S_k^S(t) := \mathcal{S}_k^S(1,t), k \in \mathbb{K}$. The following balance equations hold for each $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $t \geqslant 0$,

$$\left\langle 1, \eta_j^D(0) \right\rangle + A_j^D(t) = \left\langle 1, \eta_j^D(t) \right\rangle + S_j^D(t), \tag{9}$$

and

$$\left\langle 1, \eta_k^S(0) \right\rangle + A_k^S(t) = \left\langle 1, \eta_k^S(t) \right\rangle + S_k^D(t). \tag{10}$$

The dynamic evolution of the potential queue measures are characterized using $\boldsymbol{\mathcal{S}}^D$ and $\boldsymbol{\mathcal{S}}^S$, as shown in the following lemma, whose validity follows by [29, Theorem 2.1].

**Lemma 1.** *For each $j \in \mathbb{J}$, $k \in \mathbb{K}$, $\phi \in \mathcal{C}_c^{1,1}([0, H_j^D) \times \mathbb{R}_+)$, $\psi \in \mathcal{C}_c^{1,1}([0, H_k^S) \times \mathbb{R}_+)$, $f \in \mathcal{C}_c^1([0, H_j^D))$, $\zeta \in \mathcal{C}_c^1([0, H_k^S))$, and $t \geqslant 0$,*

$$\begin{aligned}\left\langle \phi(\cdot,t), \eta_j^D(t) \right\rangle = \left\langle \phi(\cdot,t), \eta_j^D(0) \right\rangle + \int_0^t \left\langle \phi_x(\cdot,u) + \phi_t(\cdot,u), \eta_j^D(u) \right\rangle du - \mathcal{S}_j^D(\phi,t) \\ + \int_0^t \phi(0,u) dA_j^D(u),\end{aligned} \tag{11}$$

$$\begin{aligned}\left\langle \psi(\cdot,t), \eta_k^S(t) \right\rangle = \left\langle \psi(\cdot,t), \eta_k^S(0) \right\rangle + \int_0^t \left\langle \psi_x(\cdot,u) + \psi_t(\cdot,u), \eta_k^S(u) \right\rangle du - \mathcal{S}_k^S(\psi,t) \\ + \int_0^t \psi(0,u) dA_k^S(u)\end{aligned} \tag{12}$$

$$\left\langle f, \eta_j^D(t) \right\rangle = \left\langle f, \eta_j^D(0) \right\rangle + \int_0^t \left\langle f', \eta_j^D(u) \right\rangle du - \mathcal{S}_j^D(f,t) + f(0) A_j^D(t), \tag{13}$$

$$\left\langle \zeta, \eta_k^S(t) \right\rangle = \left\langle \zeta, \eta_k^S(0) \right\rangle + \int_0^t \left\langle \zeta', \eta_k^S(u) \right\rangle du - \mathcal{S}_k^S(\zeta,t) + \zeta(0) A_k^S(t). \tag{14}$$

### 2.2.3 Matching processes

For $j \in \mathbb{J}$, $k \in \mathbb{K}$, $l = \{-\langle 1, \eta_j^D(0) \rangle + 1, \ldots, 0\} \cup \mathbb{N}$, and $h = \{-\langle 1, \eta_k^S(0) \rangle + 1, \ldots, 0\} \cup \mathbb{N}$, let $m_{jklh}$ denote the matching time of the $l$th type $j$ customer with the $h$th type $k$ worker. We set $m_{jklh} = \infty$ if the $l$th type $j$ customer and the $h$th type $k$ worker are not matched. Then, $m_{jklh}$ may be finite or infinite for $(j,k) \in \mathcal{E}$, and $m_{jklh}$ is infinite for all $(j,k) \notin \mathcal{E}$. For fixed $j$ and $l$ (resp. $k$ and $h$), $m_{jklh}$ is finite only for at most one pair $(k,h)$ (resp. $(j,l)$); i.e., one customer (resp. worker) can be matched with at most one worker (resp. customer). Specifically, we require the following inequalities:

$$\sum_{k \in \mathbb{K}} \sum_{h = -\left\langle 1, \eta_k^S(0) \right\rangle + 1}^{\infty} 1_{\{m_{jklh} < \infty\}} \leqslant 1 \text{ for each } j \in \mathbb{J} \text{ and } l \in \{-\langle 1, \eta_j^D(0) \rangle + 1, \ldots, 0\} \cup \mathbb{N},$$

and

$$\sum_{j \in \mathbb{J}} \sum_{l = -\left\langle 1, \eta_j^D(0) \right\rangle + 1}^{\infty} 1_{\{m_{jklh} < \infty\}} \leqslant 1 \text{ for each } k \in \mathbb{K} \text{ and } h \in \{-\langle 1, \eta_k^S(0) \rangle + 1, \ldots, 0\} \cup \mathbb{N}.$$

We assume that for each $j \in \mathbb{J}, k \in \mathbb{K}, l \in \{-\langle 1, \eta_j^D(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$, and $h \in \{-\langle 1, \eta_k^S(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$, if $m_{jklh} < \infty$, then

$$0 < m_{jklh} < \min\left(e_{jl}^D + r_{jl}^D, e_{kh}^S + r_{kh}^S\right), \quad \text{if } l \leqslant 0 \text{ and } h \leqslant 0, \tag{15}$$

$$\max\left(e_{jl}^D, e_{kh}^S\right) \leqslant m_{jklh} < \min\left(e_{jl}^D + r_{jl}^D, e_{kh}^S + r_{kh}^S\right), \quad \text{if either } l > 0 \text{ or } h > 0. \tag{16}$$

The inequalities in (15) enforce that a customer and a worker who are both in system at time zero can only be matched after time zero and before either of them reneges. The inequalities in (16) enforce that a customer and a worker, at least one of which arrived after time zero, can only be matched once both have arrived to the system and strictly before either of them reneges the system.

For each $j \in \mathbb{J}$ and $l \in \mathbb{N}$, the matching time of the $l$th type $j$ customer can be expressed as follows,

$$m_{jl}^D := \begin{cases} m_{jklh}, & \text{if } k \in \mathbb{K} \text{ and } h \in \{-\langle 1, \eta_k^S(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N} \text{ are such that } \ m_{jklh} < \infty, \\ \infty, & \text{otherwise.} \end{cases},$$

Similarly, for each $k \in \mathbb{K}$ and $h \in \mathbb{N}$, the matching time of the $h$th type $k$ worker can be written as follows,

$$m_{kh}^S := \begin{cases} m_{jklh}, & \text{if } \ j \in \mathbb{J} \text{ and } l \in \{-\langle 1, \eta_j^D(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N} \text{ are such that } \ m_{jklh} < \infty, \\ \infty, & \text{otherwise.} \end{cases}.$$

We assume that matchings occur between HL customers, which requires that if $-\langle 1, \eta_j^D(0)\rangle \leqslant l_1 < l_2 < \infty$ and $-\langle 1, \eta_k^S(0)\rangle \leqslant h_1 < h_2 < \infty$, then

$$m_{jl_1} \leqslant m_{jl_2}, j \in \mathbb{J}, \text{ and } m_{kh_1} \leqslant m_{kh_2}, k \in \mathbb{K}. \tag{17}$$

A *matching process* is a $\mathbb{Z}_+^{J \times K}$ valued stochastic process $\boldsymbol{M}$ defined from the matching times defined in the previous paragraph. The components of $\boldsymbol{M}$ track the cumulative number of matches between type $j \in \mathbb{J}$ customers and type $k \in \mathbb{K}$ workers in $(0, t]$, as follows: for $j \in \mathbb{J}, k \in \mathbb{K}$, and $t \geqslant 0$,

$$M_{jk}(t) := \sum_{l=-\langle 1,\eta_j^D(0)\rangle+1}^{A_j^D(t)} \sum_{h=-\langle 1,\eta_k^S(0)\rangle+1}^{A_k^S(t)} 1_{\{m_{jklh} \leqslant t\}}. \tag{18}$$

Note that $M_{jk}(0) = 0$ for all $(j, k) \in \mathcal{E}$, and that $M_{jk}(t) = 0$ for all $t \geqslant 0$ if $(j, k) \notin \mathcal{E}$.

### 2.2.4 Reneging, queue-length, and HL waiting time processes

For each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, the cumulative number of type $j$ customers $R_j^D(t)$ and type $k$ workers $R_k^S(t)$ that renege by time $t \geqslant 0$ are given by

$$R_j^D(t) := \sum_{l=-\langle 1,\eta_j^D(0)\rangle+1}^{A_j^D(t)} \sum_{s \in [0,t]} 1_{\{s \leqslant m_{jl}^D, \frac{dw_{jl}^D}{dt}(s-) > 0, \frac{dw_{jl}^D}{dt}(s+) = 0\}} \tag{19}$$

and

$$R_k^S(t) := \sum_{h=-\langle 1,\eta_k^S(0)\rangle+1}^{A_k^S(t)} \sum_{s \in [0,t]} 1_{\{s \leqslant m_{kh}^S, \frac{dw_{kh}^S}{dt}(s-) > 0, \frac{dw_{kh}^S}{dt}(s+) = 0\}}. \tag{20}$$

Note that type $j \in \mathbb{J}$ customers and type $k \in \mathbb{K}$ workers cannot be matched at the exact moment their patience time expires.

The demand and supply queue lengths for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$ at time $t \geqslant 0$ are given respectively by

$$Q_j^D(t) := Q_j^D(0) + A_j^D(t) - R_j^D(t) - \sum_{k=1}^{K} M_{jk}(t) \tag{21}$$

and

$$Q_k^S(t) := Q_k^S(0) + A_k^S(t) - R_k^S(t) - \sum_{j=1}^{J} M_{jk}(t). \tag{22}$$

The restrictions on the matching process in the previous subsection ensure that $Q_j^D(t) \geqslant 0$ and $Q_k^S(t) \geqslant 0$ for all $j \in \mathbb{J}, k \in \mathbb{K}$, and $t \geqslant 0$.

For each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, the waiting times of the HL customer and worker at time $t \geqslant 0$ are

$$\chi_j^D(t) := \inf\left\{x \in \mathbb{R}_+ : \left\langle 1_{[0,x]}, \eta_j^D(t)\right\rangle \geqslant Q_j^D(t)\right\},$$

and

$$\chi_k^S(t) := \inf\left\{x \in \mathbb{R}_+ : \left\langle 1_{[0,x]}, \eta_k^S(t)\right\rangle \geqslant Q_k^S(t)\right\},$$

respectively. Then, for each $j \in \mathbb{J}$ and $t \geqslant 0$ such that $Q_j^D(t) > 0$, it follows

$$\left\langle 1_{[0,\chi_j^D(t))}, \eta_j^D(t)\right\rangle < Q_j^D(t) \leqslant \left\langle 1_{[0,\chi_j^D(t)]}, \eta_j^D(t)\right\rangle,$$

and for each $k \in \mathbb{K}$ and $t \geqslant 0$ such that $Q_k^S(t) > 0$,

$$\left\langle 1_{[0,\chi_k^S(t))}, \eta_k^S(t)\right\rangle < Q_k^S(t) \leqslant \left\langle 1_{[0,\chi_k^S(t)]}, \eta_k^S(t)\right\rangle.$$

The HL assumption (17) implies that any type $j \in \mathbb{J}$ customer (type $k \in \mathbb{K}$ worker) waiting in queue at time $t \geqslant 0$ has been waiting in the potential queue for less than or equal to $\chi_j^D(t)$ whereas any type $j \in \mathbb{J}$ customer (type $k \in \mathbb{K}$ worker) in the potential queue at time $t \geqslant 0$ with potential waiting time strictly greater than $\chi_j^D(t)$ ($\chi_k^S(t)$) has been matched. Then, $\chi_j^D(t)$ and $\chi_k^S(t)$ are moving boundaries marking the waiting time at which potential customers and workers transition from those in queue to those not in queue because they have been matched. Note that there can be up to $\left\langle 1_{[0,\chi_j^D(t)]}, \eta_j^D(t)\right\rangle - \left\langle 1_{[0,\chi_j^D(t))}, \eta_j^D(t)\right\rangle$ type $j \in \mathbb{J}$ customers in queue at time $t \geqslant 0$ that have been waiting for time $\chi_j^D(t)$; $Q_j^D(t) - \left\langle 1_{[0,\chi_j^D(t))}, \eta_j^D(t)\right\rangle$ are in queue, and $\left\langle 1_{[0,\chi_j^D(t)]}, \eta_j^D(t)\right\rangle - Q_j^D(t)$ were matched before reneging. A similar statement holds for type $k \in \mathbb{K}$ workers in queue at time $t \geqslant 0$.

Similar to [43, Inequalities 29 and 30], this implies the following upper and lower bounds for the number of reneging customers and workers at time $t \geqslant 0$:

$$R_j^D(t) \leqslant \sum_{l=-\left\langle 1,\eta_j^D(0)\right\rangle+1}^{A_j^D(t)} \sum_{s\in(0,t]} 1_{\left\{s\leqslant\chi_j^D(s-),\, \frac{dw_{jl}^D}{dt}(s-)>0,\, \frac{dw_{jl}^D}{dt}(s+)=0\right\}}, \tag{23}$$

$$R_j^D(t) \geqslant \sum_{l=-\left\langle 1,\eta_j^D(0)\right\rangle+1}^{A_j^D(t)} \sum_{s\in(0,t]} 1_{\left\{s<\chi_j^D(s-),\, \frac{dw_{jl}^D}{dt}(s-)>0,\, \frac{dw_{jl}^D}{dt}(s+)=0\right\}}, \tag{24}$$

for each $j \in \mathbb{J}$, and

$$R_k^S(t) \leqslant \sum_{h=-\langle 1,\eta_k^S(0)\rangle+1}^{A_k^S(t)} \sum_{s\in(0,t]} 1_{\{s\leqslant \chi_k^S(s-),\frac{dw_{kh}^S}{dt}(s-)>0,\frac{dw_{kh}^S}{dt}(s+)=0\}}, \tag{25}$$

$$R_k^S(t) \geqslant \sum_{h=-\langle 1,\eta_k^S(0)\rangle+1}^{A_k^S(t)} \sum_{s\in(0,t]} 1_{\{s< \chi_k^S(s-),\frac{dw_{kh}^S}{dt}(s-)>0,\frac{dw_{kh}^S}{dt}(s+)=0\}}, \tag{26}$$

for each $k \in \mathbb{K}$. The bound on the left-hand side in (23) includes customers that have waited the same amount of time as the HL customer (because $s \leqslant \chi_j^D(s-)$ in the indicator function) whereas the bound on the left-hand side in (24) only includes customers that have waited strictly less than the HL customer (because $s < \chi_j^D(s-)$ in the indicator function), and similar holds true for the bounds in (25) and (26). Note that if the arrival processes have jumps of size one (meaning customers and workers do not arrive in batches), then (23) and (25) hold with equality.

### 2.3 Admissible matching policies

The dynamic equations and conditions specified in Section 2.2 on the matching process $\boldsymbol{M}$ are fundamental for an HL-matching model. For the analysis here, we consider matching processes that render matching decisions based on past and current information, i.e., do not use information about the future. For this, we note that the matching process may be such that the states that can be achieved live in a strict subset of $\mathbb{Y}$. For example, if the matching policy prioritizes matches between type $j \in \mathbb{J}$ customers and type $k \in \mathbb{K}$ workers, then the matching policy will disallow states in which both customer type $j$ and worker type $k$ are present, i.e., $Q_j^D(t)Q_k^S(t) = 0$ for all $t \geqslant 0$. To account for this, we introduce a subspace $\mathbb{X}$ of $\mathbb{Y}$ in the next definition.

**Definition 1.** *A **matching policy** is a pair $(\mathbb{X}, \{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\})$ where $\mathbb{X}$ is a Polish subspace of $\mathbb{Y}$ and $\{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\}$ is a collection of probability measures on $(\Omega, \mathcal{F})$ such that the following hold:*

1. *For each $\boldsymbol{y} \in \mathbb{X}$, $\mathbb{P}_{\boldsymbol{y}}(\boldsymbol{Y} \in \mathcal{D}(\mathbb{X}), \boldsymbol{Y}(0) = \boldsymbol{y}$ and $\boldsymbol{Y}$ satisfies (3) − (26)$) = 1$;*
2. *For any measurable $B \subseteq \mathcal{D}(\mathbb{X})$, the mapping $\boldsymbol{y} \to \mathbb{P}_{\boldsymbol{y}}(\boldsymbol{Y} \in B)$ from $\mathbb{X}$ to $[0,1]$ is Borel measurable.*

Given a matching policy $(\mathbb{X}, \{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\})$ and $\boldsymbol{y} \in \mathbb{X}$, we let $\mathcal{L}_{\boldsymbol{y}}$ denote the law of the state process $\boldsymbol{Y}$ with $\boldsymbol{Y}(0) = \boldsymbol{y}$, i.e., $\mathcal{L}_{\boldsymbol{y}}(B) = \mathbb{P}_{\boldsymbol{y}}(\boldsymbol{Y} \in B)$ for all Borel measurable $B \subset \mathcal{D}(\mathbb{X})$.

For a matching policy $(\mathbb{X}, \{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\})$ to be admissible, we require the associated matching process to be nonanticipating in the sense that it is adapted to the filtration determined by the history of the state process, which we define precisely here. Given a matching policy $(\mathbb{X}, \{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\})$ and $\boldsymbol{y} \in \mathbb{X}$, we let $\boldsymbol{Y} = \left(\boldsymbol{a}^D, \boldsymbol{a}^S, \boldsymbol{Q}^D, \boldsymbol{Q}^S, \boldsymbol{\eta}^D, \boldsymbol{\eta}^S\right)$ denote the state process that has law $\mathcal{L}_{\boldsymbol{y}}$ and define

$$\tilde{\boldsymbol{Y}}(t) := \left(\boldsymbol{a}^D(t), \boldsymbol{a}^S(t), \boldsymbol{Q}^D(t-), \boldsymbol{Q}^S(t-), \boldsymbol{\eta}^D(t), \boldsymbol{\eta}^S(t)\right), \qquad \text{for each } t \geqslant 0, \tag{27}$$

where $\left(\boldsymbol{Q}^D(0-), \boldsymbol{Q}^S(0-)\right) = \left(\boldsymbol{Q}^D(0), \boldsymbol{Q}^S(0)\right)$. We further define the filtration $\{\mathcal{G}_t^{\boldsymbol{y}}\}_{t\geqslant 0}$ such that

$$\mathcal{G}_t^{\boldsymbol{y}} = \sigma\left(\left\{\tilde{\boldsymbol{Y}}(s), 0 \leqslant s \leqslant t\right\}\right), \qquad \text{for each } t \geqslant 0.$$

Then, for each $t \geqslant 0$, the $\sigma$-algebra $\mathcal{G}_t^{\boldsymbol{y}}$ includes information about customers and workers waiting to be matched immediately before time $t$, and customers and workers arriving or reneging (actual or virtual) at time $t$. This is the information that should naturally be used to determine if and which customers and workers to match at time $t$.

**Definition 2.** *A matching policy* $(\mathbb{X}, \{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\})$ *is said to be* ***admissible*** *if for each* $\boldsymbol{y} \in \mathbb{X}$*, the matching process* $\boldsymbol{M}$ *given in* (18) *for the state process with* $\boldsymbol{Y}$ *with law* $\mathcal{L}_{\boldsymbol{y}}$ *is* $\left\{\mathcal{G}_t^{\boldsymbol{y}}\right\}_{t\geqslant 0}$*-adapted.*

## 2.4 Initial Conditions

Here we introduce random initial conditions. For this we recall that all random elements are defined on the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $(\mathbb{X}, \{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\})$ be a matching policy and define

$$\Xi_0 := \{\boldsymbol{Y}_0 : \mathbb{P}(\boldsymbol{Y}_0 \in \mathbb{X}) = 1 \text{ and } \boldsymbol{Y}_0 \text{ is independent of the stochastic primitives}\}.$$

Note that $\Xi_0$ depends on the matching policy through its dependence on the subspace $\mathbb{X}$. Given $\boldsymbol{Y}_0 \in \Xi_0$, we let $\xi$ denote the law of $\boldsymbol{Y}_0$, i.e., $\xi(B) = \mathbb{P}(\boldsymbol{Y}_0 \in B)$ for all Borel measurable $B \subset \mathbb{X}$, and, with a slight abuse of notation, we will sometimes write $\xi \in \Xi_0$. Given $\xi \in \Xi_0$, we define the Borel probability measure $\mathcal{L}_\xi$ on $\mathcal{D}(\mathbb{X})$ such that for each Borel measurable $B \subseteq \mathcal{D}(\mathbb{X})$ satisfies

$$\mathcal{L}_\xi(B) = \int_{\mathbb{X}} \mathcal{L}_{\boldsymbol{y}}(B)\xi(d\boldsymbol{y}). \tag{28}$$

Then $\mathcal{L}_\xi$ denotes the law of the state process such the initial condition has distribution $\xi$. We let $\mathbb{E}_\xi$ denote the expectation operator with respect to $\mathcal{L}_\xi$. In our analysis, we restrict attention to random initial conditions for which the expected number of potential customers and potential workers in system at time zero are finite. In particular, we restrict attention to $\xi \in \Xi_0$ such that

$$\max_{j\in\mathbb{J}} \mathbb{E}_\xi\left[\langle 1, \eta_j^D(0)\rangle\right] < \infty \qquad \text{and} \qquad \max_{k\in\mathbb{K}} \mathbb{E}_\xi\left[\langle 1, \eta_k^S(0)\rangle\right] < \infty. \tag{29}$$

We also restrict attention to initial conditions such that the expected number of exogenous arrivals in to the system in the time interval $(0, t]$ is finite for all $t \geqslant 0$. That is, we restrict attention to $\xi \in \Xi_0$ such that for all $t \geqslant 0$,

$$\max_{j\in\mathbb{J}} \mathbb{E}_\xi\left[A_j^D(t)\right] < \infty \qquad \text{and} \qquad \max_{k\in\mathbb{K}} \mathbb{E}_\xi\left[A_k^S(t)\right] < \infty. \tag{30}$$

Let

$$\Xi := \{\xi \in \Xi_0 : (29) \text{ and } (30) \text{ hold}\}.$$

Due to (1), $\delta_{\boldsymbol{y}} \in \Xi$ for all $\boldsymbol{y} \in \mathbb{X}$.

For $\xi \in \Xi$, due to (5), (6), (28), (29) and (30), the processes in (7) and (8) that arise when starting from the initial condition $\xi$ are bounded in expectation for every $t \geqslant 0$ when $\phi$ and $\psi$ are bounded functions. In particular, for all $\xi \in \Xi$, $j \in \mathbb{J}, t \geqslant 0$, and bounded measurable $\phi : \left[0, H_j^D\right) \to \mathbb{R}$,

$$\mathbb{E}_\xi\left[|\mathcal{S}_j^D(\phi, t)|\right] \leqslant \|\phi\|_\infty \mathbb{E}_\xi\left[\left\langle 1, \eta_j^D(0)\right\rangle + A_j^D(t)\right] < \infty$$

and for all $k \in \mathbb{K}, t \geqslant 0$, and bounded measurable $\psi : \left[0, H_k^S\right) \to \mathbb{R}$,

$$\mathbb{E}_\xi\left[|\mathcal{S}_k^S(\psi, t)|\right] \leqslant \|\psi\|_\infty \mathbb{E}_\xi\left[\left\langle 1, \eta_k^S(0)\right\rangle + A_k^S(t)\right] < \infty.$$

Furthermore, the following lemma presents dynamic evolution equations in expectation, and it is a consequence of [30, Proposition 2.2], noting the finiteness conditions in (29) and (30).

**Lemma 2.** *Suppose that $\xi \in \Xi$. For $j \in \mathbb{J}$, any bounded measurable function $f : [0, H_j^D) \to \mathbb{R}$, and $t \geqslant 0$,*

$$\mathbb{E}_\xi\left[\left\langle f, \eta_j^D(t)\right\rangle\right] = \mathbb{E}_\xi\left[\int_0^{H_j^D} f(x+t)\frac{1-G_j^D(x+t)}{1-G_j^D(x)}\eta_j^D(0)(dx)\right] + \mathbb{E}_\xi\left[\int_0^t f(t-u)(1-G_j^D(t-u))dA_j^D(u)\right],$$

*and, for $k \in \mathbb{K}$, any bounded Borel measurable function $f : [0, H_j^S) \to \mathbb{R}$ and $t \geqslant 0$,*

$$\mathbb{E}_\xi\left[\left\langle f, \eta_k^S(t)\right\rangle\right] = \mathbb{E}_\xi\left[\int_0^{H_k^S} f(x+t)\frac{1-G_k^S(x+t)}{1-G_k^S(x)}\eta_k^S(0)(dx)\right] + \mathbb{E}_\xi\left[\int_0^t f(t-u)(1-G_k^S(t-u))dA_k^S(u)\right].$$

# 3 A fluid model

In this section, we present a fluid model which can be seen as an approximation of the stochastic model introduced in Section 2. Section 3.1 provides the fluid model equations and defines a fluid model solution. Section 3.2 provides conditions for uniqueness of fluid model solutions.

## 3.1 Fluid model solutions

Recall that $H_j^D$ and $H_k^S$ are the right edges of the support of the cumulative reneging distribution functions for any $j \in \mathbb{J}$ and $k \in \mathbb{K}$. A fluid model solution takes values in

$$\overline{\mathbb{Y}}_0 := \mathbb{R}_+^J \times \mathbb{R}_+^K \times \left(\times_{j=1}^J \mathcal{M}[0, H_j^D)\right) \times \left(\times_{j=1}^K \mathcal{M}[0, H_j^S)\right).$$

Roughly speaking, a fluid model solution is represented by a vector $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S) \in \mathcal{C}(\overline{\mathbb{Y}}_0)$ that is the analogue of the state descriptor. The first functions $\overline{\boldsymbol{Q}}^D$ and $\overline{\boldsymbol{Q}}^S$ represent the fluid queue lengths and the measure-valued functions $\overline{\boldsymbol{\eta}}^D$ and $\overline{\boldsymbol{\eta}}^S$ represent the fluid potential queue measures (the analogues of (3) and (4)). We consider a subset $\overline{\mathbb{Y}}$ of $\overline{\mathbb{Y}}_0$ in which, analogous to (2), the fluid queue lengths cannot exceed the fluid potential queues. Specifically, if $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S) \in \mathcal{C}(\overline{\mathbb{Y}})$, then for all $t \geqslant 0$

$$\overline{Q}_j^D(t) \leqslant \left\langle 1, \overline{\eta}_j^D(t)\right\rangle, \text{ for all } j \in \mathbb{J}, \text{ and } \overline{Q}_k^S(t) \leqslant \left\langle 1, \overline{\eta}_k^S(t)\right\rangle, \text{ for all } k \in \mathbb{K}.$$

A fluid model solution $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S) \in \mathcal{C}(\overline{\mathbb{Y}})$ satisfies finiteness conditions such that for all $t \geqslant 0$

$$\int_0^t \left\langle h_j^D, \overline{\eta}_j^D(u)\right\rangle du < \infty, \text{ for all } j \in \mathbb{J}, \text{ and } \int_0^t \left\langle h_k^S, \overline{\eta}_k^S(u)\right\rangle du < \infty, \text{ for all } k \in \mathbb{K}, \tag{31}$$

and has initial potential queue measures with no atoms; i.e.,

$$\left\langle 1_{\{x\}}, \overline{\eta}_j^D(0)\right\rangle = 0 \text{ for all } x \in [0, H_j^D), j \in \mathbb{J} \text{ and } \left\langle 1_{\{x\}}, \overline{\eta}_k^S(0)\right\rangle = 0 \text{ for all } x \in [0, H_K^S), k \in \mathbb{K}. \tag{32}$$

The fluid analogues of the cumulative reneging processes (19) and (20) are, for $t \geqslant 0$ and each $j \in \mathbb{J}$ and $k \in \mathbb{K}$,

$$\overline{R}_j^D(t) = \int_0^t \int_0^{\overline{Q}_j^D(u)} h_j^D\Big((F_{j,u}^D)^{-1}(y)\Big)dydu, \text{ and } \overline{R}_k^S(t) = \int_0^t \int_0^{\overline{Q}_k^S(u)} h_k^S\Big((F_{k,u}^S)^{-1}(y)\Big)dydu, \tag{33}$$

where for each $j \in \mathbb{J}$, $k \in \mathbb{K}$, $x \in \mathbb{R}_+$, $y \in \mathbb{R}_+$, and $u \geqslant 0$, we define

$$F_{j,u}^D(x) := \left\langle 1_{[0,x]}, \eta_j^D(u) \right\rangle, \text{ and } F_{k,u}^S(x) := \left\langle 1_{[0,x]}, \eta_k^S(u) \right\rangle,$$

and

$$(F_{j,u}^D)^{-1}(y) := \inf\{x \in \mathbb{R}_+ : F_{j,u}^D(x) \geqslant y\}, \text{ and } (F_{k,u}^S)^{-1}(y) := \inf\{x \in \mathbb{R}_+ : F_{k,u}^S(x) \geqslant y\},$$

noting that $\inf \varnothing = \infty$. The condition (31) ensures that the cumulative amount of fluid reneging in (33) is finite for all time.

The input to the fluid model are componentwise non-decreasing functions $\overline{\boldsymbol{A}}^D \in \mathcal{C}(\mathbb{R}_+^J)$ and $\overline{\boldsymbol{A}}^S \in \mathcal{C}(\mathbb{R}_+^K)$ with $\overline{\boldsymbol{A}}^D(0) = \mathbf{0}$ and $\overline{\boldsymbol{A}}^S(0) = \mathbf{0}$ that we term arrival functions. For any continuous and bounded function $f \in \mathcal{C}_b(\mathbb{R}_+)$ the following integral equations hold for each $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $t \geqslant 0$,

$$\left\langle f, \overline{\eta}_j^D(t) \right\rangle = \int_0^{H_j^D} f(x+t)\frac{1-G_j^D(x+t)}{1-G_j^D(x)}\overline{\eta}_j^D(0)(dx) + \int_0^t f(t-u)(1-G_j^D(t-u))d\overline{A}_j^D(u), \tag{34}$$

and

$$\left\langle f, \overline{\eta}_k^S(t) \right\rangle = \int_0^{H_k^S} f(x+t)\frac{1-G_k^S(x+t)}{1-G_k^S(x)}\overline{\eta}_k^S(0)(dx) + \int_0^t f(t-u)(1-G_k^S(t-u))d\overline{A}_k^S(u). \tag{35}$$

The equations (34) and (35) parallel the dynamic evolution equations presented for the stochastic model in Lemma 2 (see equations (11)-(14) and marked point process definitions (7) and (8)). Instead of (34) and (35), we could use similar integral equations as in (11) and (12) in Lemma 1. However, both are equivalent as we state in the following remark.

**Remark 1.** *By [32, Theorem 4.1] (see also [43, Remark 1]), if $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ are arrival functions and $(\overline{\boldsymbol{\eta}}^{\boldsymbol{D}}, \overline{\boldsymbol{\eta}}^{\boldsymbol{S}})$ satisfies (31), then (34) and (35) hold if and only if the following hold: for all $j \in \mathbb{J}$, $k \in \mathbb{K}$, $\phi \in \mathcal{C}_c^{1,1}([0, H_j^D) \times \mathbb{R}_+)$, $\psi \in \mathcal{C}_c^{1,1}([0, H_k^S) \times \mathbb{R}_+)$, and $t \geqslant 0$,*

$$\begin{aligned}
\left\langle \phi(\cdot,t), \overline{\eta}_j^D(t) \right\rangle &= \left\langle \phi(\cdot,0), \overline{\eta}_j^D(0) \right\rangle + \int_0^t \left\langle \phi_x(\cdot,u) + \phi_t(\cdot,u), \overline{\eta}_j^D(u) \right\rangle du \\
&\quad - \int_0^t \left\langle h_j^s(\cdot)\phi(\cdot,u), \overline{\eta}_j^D(u) \right\rangle du + \int_0^t \phi(0,u)d\overline{A}_j^D(u), \\
\left\langle \psi(\cdot,t), \overline{\eta}_k^S(t) \right\rangle &= \left\langle \psi(\cdot,0), \overline{\eta}_k^D(0) \right\rangle + \int_0^t \left\langle \psi_x(\cdot,u) + \psi_t(\cdot,u), \overline{\eta}_k^S(u) \right\rangle du \\
&\quad - \int_0^t \left\langle h_k^S(\cdot)\psi(\cdot,u), \overline{\eta}_k^S(u) \right\rangle du + \int_0^t \psi(0,u)d\overline{A}_k^S(u).
\end{aligned}$$

The specification of a fluid model solution for given arrival functions $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ and an initial condition requires the specification of a matching function. The matching function can be thought of as the fluid analogue to the matching policy for the stochastic system given in (18), and satisfying restrictions (16) and (17).

**Definition 3.** *A* ***matching function*** $\overline{\boldsymbol{M}} \in \mathcal{C}(\mathbb{R}_+^{J\times K})$ *is a matrix of componentwise non-decreasing functions such that* $\overline{M}_{jk}(0) = 0$ *for all* $j \in \mathbb{J}$ *and* $k \in \mathbb{K}$.

For a given matching function $\overline{\boldsymbol{M}}$, $\overline{M}_{jk}(t)$ is interpreted as the amount of type $j$ customer fluid and type $k$ worker fluid matched by time $t$, for $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $t \geqslant 0$. Then, the fluid queue-lengths evolve as follows: for all $j \in J$, $k \in \mathbb{K}$ and $t \geqslant 0$,

$$\overline{Q}_j^D(t) = \overline{Q}_j^D(0) + \overline{A}_j^D(t) - \overline{R}_j^D(t) - \sum_{k\in\mathbb{K}} \overline{M}_{jk}(t), \tag{36}$$

and,

$$\overline{Q}_k^S(t) = \overline{Q}_k^S(0) + \overline{A}_k^S(t) - \overline{R}_k^S(t) - \sum_{j\in\mathbb{J}} \overline{M}_{jk}(t). \tag{37}$$

The equations (36) and (37) are the fluid analogues of the queue-length evolution equations (21) and (22) in the stochastic model.

**Definition 4.** *Let* $\overline{\boldsymbol{A}}^D$ *and* $\overline{\boldsymbol{A}}^S$ *be arrival functions. A* ***fluid model solution*** *for* $(\overline{\boldsymbol{A}}^D, \overline{\boldsymbol{A}}^S)$ *is* $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S) \in \mathcal{C}(\overline{\mathbb{Y}})$ *that satisfies conditions* (31) *and* (32)*, the integral equations* (34) *and* (35)*, and is such that there exists a matching function* $\overline{\boldsymbol{M}}$ *for which* (36) *and* (37) *hold, with* $\overline{\boldsymbol{R}}^D$ *and* $\overline{\boldsymbol{R}}^S$ *given by* (33).

There is an alternative, potentially more intuitive, representation of the reneging process, given in the following remark.

**Remark 2.** *Suppose that* $\overline{\boldsymbol{A}}^D$ *and* $\overline{\boldsymbol{A}}^S$ *are arrival functions and* $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S) \in \mathcal{C}(\overline{\mathbb{Y}})$ *satisfies (31) and (32). Then,* $\overline{\boldsymbol{\eta}}_j^D(t)$ *and* $\overline{\boldsymbol{\eta}}_k^S(t)$ *have no atoms for all* $j \in \mathbb{J}$, $k \in \mathbb{K}$, $t \geqslant 0$, *and the following equations hold: for all* $j \in \mathbb{J}, k \in \mathbb{K}$, *and* $t \geqslant 0$,

$$\overline{R}_j^D(t) = \int_0^t \int_0^{H_j^D} h_j^D(x) 1_{\{\overline{\eta}_j^D(u)[0,x] < \overline{Q}_j^D(u)\}} \overline{\eta}_j^D(u)(dx)du \tag{38}$$

*and*

$$\overline{R}_k^S(t) = \int_0^t \int_0^{H_k^S} h_k^S(x) 1_{\{\overline{\eta}_k^S(u)[0,x] < \overline{Q}_k^S(u)\}} \overline{\eta}_k^S(u)(dx)du. \tag{39}$$

The inner integrals in (38) and (39) represent the instantaneous reneging rate, which is determined by the hazard rate function and fluid age. Then, integrating over the instantaneous reneging rate in $[0, t]$ gives the cumulative reneging up to time $t$.

A fluid model solution for arrival functions $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ that arise as functional law of large number limits of the arrival processes $\boldsymbol{A}^D$ and $\boldsymbol{A}^S$ in the stochastic model provides an approximation for the mean queue-lengths at each time $t \geqslant 0$. Fluid model solutions are more tractable than the original stochastic model, yet they are still somewhat complicated because they involve measure-valued functions. Even so, one could apply numerical methods to find the fluid model solutions as in [37] and [38].

### 3.2 Existence and uniqueness

A fundamental question is if a solution of the fluid model exists and if it is unique. Our first result provides conditions for uniqueness when a fluid model solution exists.

**Theorem 1.** *Let $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ be arrival functions and let $\overline{\boldsymbol{M}}$ be a matching function. Suppose $(\overline{\boldsymbol{Q}}^{D,1}, \overline{\boldsymbol{Q}}^{S,1}, \overline{\boldsymbol{\eta}}^{D,1}, \overline{\boldsymbol{\eta}}^{S,1})$ and $(\overline{\boldsymbol{Q}}^{D,2}, \overline{\boldsymbol{Q}}^{S,2}, \overline{\boldsymbol{\eta}}^{D,2}, \overline{\boldsymbol{\eta}}^{S,2})$ are both fluid model solutions for $\left(\overline{\boldsymbol{A}}^D, \overline{\boldsymbol{A}}^S\right)$ that satisfy (36) and (37) for the matching function $\overline{\boldsymbol{M}}$, and*

$$(\overline{\boldsymbol{Q}}^{D,1}(0), \overline{\boldsymbol{Q}}^{S,1}(0), \overline{\boldsymbol{\eta}}^{D,1}(0), \overline{\boldsymbol{\eta}}^{S,1}(0)) = (\overline{\boldsymbol{Q}}^{D,2}(0), \overline{\boldsymbol{Q}}^{S,2}(0), \overline{\boldsymbol{\eta}}^{D,2}(0), \overline{\boldsymbol{\eta}}^{S,2}(0)).$$

*Then,*

$$(\overline{\boldsymbol{Q}}^{D,1}, \overline{\boldsymbol{Q}}^{S,1}, \overline{\boldsymbol{\eta}}^{D,1}, \overline{\boldsymbol{\eta}}^{S,1}) = (\overline{\boldsymbol{Q}}^{D,2}, \overline{\boldsymbol{Q}}^{S,2}, \overline{\boldsymbol{\eta}}^{D,2}, \overline{\boldsymbol{\eta}}^{S,2})$$

***Proof.*** By (34) and (35), we directly have that $\overline{\boldsymbol{\eta}}^{D,1} = \overline{\boldsymbol{\eta}}^{D,2}$ and $\overline{\boldsymbol{\eta}}^{S,1} = \overline{\boldsymbol{\eta}}^{S,2}$, and so, to ease the notation in what follows, we define $(F^D_{j,t})^{-1}$ and $(F^S_{k,t})^{-1}$ as in the display following (33) for all $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $t \geqslant 0$, without adding superscripts 1 and 2.

We shall argue by contradiction that the fluid queue lengths are also identical. Fix a $j \in \mathbb{J}$. Assume that there exists $t^\star$ such that $\overline{Q}^{D,1}_j(t^\star) > \overline{Q}^{D,2}_j(t^\star)$ and define $u = \sup\{0 \leqslant s < t^\star : \overline{Q}^{D,1}_j(s) \leqslant \overline{Q}^{D,2}_j(s)\} \vee 0$, where by convention $\sup \varnothing = -\infty$. Due to continuity, $\overline{Q}^{D,1}_j(u) = \overline{Q}^{D,2}_j(u)$ and $\overline{Q}^{D,1}_j(s) > \overline{Q}^{D,2}_j(s)$ for $s \in (u, t^\star]$. If $u = 0$, then define $\overline{Q}^{D,1}_j(0-) = \overline{Q}^{D,1}_j(0)$ and the same hold for all the functions. From (33), we have that

$$\begin{aligned}
\overline{R}^{D,1}_j(t^\star) - \overline{R}^{D,1}_j(u) &= \int_u^{t^\star} \int_0^{\overline{Q}^{D,1}_j(s)} h^D_j\left((F^D_{j,s})^{-1}(y)\right) dyds \\
&\geqslant \int_u^{t^\star} \int_0^{\overline{Q}^{D,2}_j(s)} h^D_j\left((F^D_{j,s})^{-1}(y)\right) dyds \\
&= \overline{R}^{D,2}_j(t^\star) - \overline{R}^{D,2}_j(u).
\end{aligned}$$

By (36), we obtain for $s \geqslant 0$,

$$\overline{Q}^{D,1}_j(s) + \overline{R}^{D,1}_j(s) = \overline{Q}^{D,2}_j(s) + \overline{R}^{D,2}_j(s).$$

By the continuity of $\overline{R}^{D,i}_j$ and the last two relations, we have that

$$\overline{Q}^{D,1}_j(t^\star) - \overline{Q}^{D,2}_j(t^\star) = \overline{R}^{D,2}_j(t^\star) - \overline{R}^{D,1}_j(t^\star) \leqslant \overline{R}^{D,2}_j(u) - \overline{R}^{D,1}_j(u) = \overline{Q}^{D,1}_j(u) - \overline{Q}^{D,2}_j(u) = 0.$$

That is,

$$\overline{Q}^{D,1}_j(t^\star) \leqslant \overline{Q}^{D,2}_j(t^\star),$$

which is a contradiction. Hence $\overline{Q}^{D,1}_j(t) \leqslant \overline{Q}^{D,2}_j(t)$ for each $t \geqslant 0$. Using exactly the symmetric arguments, we have that $\overline{Q}^{D,1}_j(t) \geqslant \overline{Q}^{D,2}_j(t)$ for each $t \geqslant 0$, and hence $\overline{Q}^{D,1}_j(t) = \overline{Q}^{D,2}_j(t)$ for each $t \geqslant 0$. The uniqueness for the fluid supply queue lengths shares the same machinery. □

Having defined the fluid model and studied its properties, we move in the next section to show how a fluid model arises. In particular, the existence of a fluid model solution follows from Theorem 3 below.

# 4 Fluid limit points

In this section, we rigorously show that a fluid model solution arises as a limit point of a sequence of fluid-scaled state descriptors. Consider a family of systems indexed by $n \in \mathbb{N}$, that are all defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and that share the patience time sequences $\{r_{jl}^D\}_{l\in\mathbb{N}}, j \in \mathbb{J}$, and $\{r_{kh}^S\}_{h\in\mathbb{N}}, k \in \mathbb{K}$. However, the arrival processes, admissible matching policies and initial conditions depend on $n \in \mathbb{N}$. Then, for each $n \in \mathbb{N}$, there are arrival processes $\boldsymbol{A}^{D,n}$ and $\boldsymbol{A}^{S,n}$, an admissible matching policy $\left(\mathbb{X}^n, \{\mathbb{P}_{\boldsymbol{y}}^n : y \in \mathbb{X}^n\}\right)$ and an initial condition $\boldsymbol{Y}^n(0) \in \Xi^n$, such that the state process $\boldsymbol{Y}^n$ with initial condition $\boldsymbol{Y}^n(0)$ has the associated matching process $\boldsymbol{M}^n$ (see Definition 2 and Section 2.4). The fluid-scaled processes for the $n$th system are as follows: for $\boldsymbol{H}^n = \boldsymbol{A}^{D,n}, \boldsymbol{A}^{S,n}, \boldsymbol{Q}^{D,n}, \boldsymbol{Q}^{S,n}, \boldsymbol{\eta}^{D,n}, \boldsymbol{\eta}^{S,n}, \boldsymbol{R}^{D,n}, \boldsymbol{R}^{S,n}, \boldsymbol{M}^n, \boldsymbol{\mathcal{S}}^{D,n}, \boldsymbol{\mathcal{S}}^{S,n}$, let $\overline{H}^n = H^n/n$. The only processes not scaled by $n$ are the processes tracking the time elapsed since the last arrival, so that $\overline{\boldsymbol{a}}^{D,n} = \boldsymbol{a}^{D,n}$ and $\overline{\boldsymbol{a}}^{S,n} = \boldsymbol{a}^{S,n}$. Then, the fluid-scaled state process is $\overline{\boldsymbol{Y}}^n = (\boldsymbol{a}^{D,n}, \boldsymbol{a}^{S,n}, \overline{\boldsymbol{Q}}^{D,n}, \overline{\boldsymbol{Q}}^{S,n}, \overline{\boldsymbol{\eta}}^{D,n}, \overline{\boldsymbol{\eta}}^{S,n})$ for $n \in \mathbb{N}$. To avoid cluttering the notation for each $n \in \mathbb{N}$, we will use $\mathbb{P}$ and $\mathbb{E}$ instead of $\mathbb{P}_{\xi^n}^n$ and $\mathbb{E}_{\xi^n}^n$ throughout.

The results proved in this section (see Theorems 2 and 3 below) hold under the assumptions stated in the paragraphs that follow. These assumptions parallel Assumptions 1-5 in [43], but are modified from their multiclass many-server queue with reneging setting to our matching setting (which involves ignoring any assumptions on their service measure, and replacing assumptions on their entry-into-service process with similar ones on our matching process). The assumptions are consistent with those required in [29], where the single class many-server queue with reneging setting is studied, with the same caveats. Another difference with the aforementioned papers is that we require our fluid model solutions to be continuous (see Definition 4), and so enforce that the family of fluid-scaled matching process is $\mathcal{C}$-tight.

The first three assumptions below are used to prove that the sequence of fluid-scaled state descriptors is tight. These assumptions ensure (1) that the arrival processes are convergent under fluid-scaling, (2) that the oscillations of the matching process can be controlled, and (3) that the initial conditions converge to a "good" state.

**Assumption 1.** *There are processes $\overline{\boldsymbol{A}}^D \in \mathcal{D}(\mathbb{R}_+^J)$ and $\overline{\boldsymbol{A}}^S \in \mathcal{D}(\mathbb{R}_+^K)$ such that for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$,*

1. *$\lim_{n\to\infty} \overline{A}_j^{D,n} = \overline{A}_j^D$ and $\lim_{n\to\infty} \overline{A}_k^{S,n} = \overline{A}_k^S$, $\mathbb{P}$-almost surely,*
2. *$\lim_{n\to\infty} \mathbb{E}\left[\overline{A}_j^{D,n}(t)\right] = \mathbb{E}\left[\overline{A}_j^D(t)\right] < \infty$ and $\lim_{n\to\infty} \mathbb{E}\left[\overline{A}_k^{D,n}(t)\right] = \mathbb{E}\left[\overline{A}_k^S(t)\right] < \infty$ for all $t \geqslant 0$.*

We remark that $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ as in Assumption 1 are necessarily componentwise non-decreasing and satisfy $\overline{A}_j^D(0) = 0$ for all $j \in \mathbb{J}$ and $\overline{A}_k^S(0) = 0$ for all $k \in \mathbb{K}$, $\mathbb{P}$-almost surely. Hence, $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ are arrival functions (as defined in Section 3) $\mathbb{P}$-almost surely.

**Assumption 2.** *We assume that for all $n \in \mathbb{N}$ either $\overline{M}_{jk}^n$ satisfies the second condition (K.2) of Kurtz' criteria for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$ or for all $0 \leqslant s \leqslant t < \infty$,*

$$\max_{j\in\mathbb{J},k\in\mathbb{K}} (\overline{M}_{jk}^n(t) - \overline{M}_{jk}^n(s)) \leqslant \sum_{j\in\mathbb{J}} (\overline{A}_j^{D,n}(t) - \overline{A}_j^{D,n}(s)) + \sum_{k\in\mathbb{K}} (\overline{A}_k^{S,n}(t) - \overline{A}_k^{S,n}(s)). \tag{40}$$

Recall **Kurtz' Criteria** for tightness (see, e.g., [20, Theorem 3.8.6 and Remark 3.8.7]). A sequence of processes $\{H^N\}_{N\in\mathbb{N}}$ with sample paths in $\mathbf{D}(\mathbb{R})$ is relatively compact if and only if the following two properties hold:

(K.1) For all rational $t \geqslant 0$, $\lim_{M\to\infty} \sup_N \mathbb{P}\left(|H^N(t)| > M\right) = 0$.

(K.2) For all rational $t > 0$, there exists $q > 0$ such that $\lim_{\varepsilon\to 0} \sup_N \mathbb{E}\left[|H^N(t+\varepsilon) - H^N(t)|^q\right] = 0$.

**Assumption 3.** *There is a random element $(\tilde{\boldsymbol{Q}}^D(0), \tilde{\boldsymbol{Q}}^S(0), \tilde{\boldsymbol{\eta}}^D(0), \tilde{\boldsymbol{\eta}}^S(0)) \in \overline{\mathbb{Y}}$, $\mathbb{P}$-almost surely, such that for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$,*

1. *$\lim_{n\to\infty} \overline{Q}_j^{D,n}(0) = \tilde{Q}_j^D(0)$ and $\lim_{n\to\infty} \overline{Q}_k^{S,n}(0) = \tilde{Q}_k^S(0)$, $\mathbb{P}$-almost surely,*
2. *$\lim_{n\to\infty} \mathbb{E}\left[\overline{Q}_j^{D,n}(0)\right] = \mathbb{E}\left[\tilde{Q}_j^D(0)\right] < \infty$ and $\lim_{n\to\infty} \mathbb{E}\left[\overline{Q}_k^{S,n}(0)\right] = \mathbb{E}\left[\tilde{Q}_k^S(0)\right] < \infty$,*
3. *$\overline{\eta}_j^{D,n}(0) \xrightarrow{w} \tilde{\eta}_j^D(0)$ and $\overline{\eta}_k^{S,n}(0) \xrightarrow{w} \tilde{\eta}_k^S(0)$, as $n \to \infty$, $\mathbb{P}$-almost surely,*
4. *$\lim_{n\to\infty} \mathbb{E}\left[\left\langle 1, \overline{\eta}_j^{D,n}(0)\right\rangle\right] = \mathbb{E}\left[\left\langle 1, \tilde{\eta}_j^D(0)\right\rangle\right] < \infty$ and $\lim_{n\to\infty} \mathbb{E}\left[\left\langle 1, \overline{\eta}_k^{S,n}(0)\right\rangle\right] = \mathbb{E}\left[\langle 1, \tilde{\eta}_k^S(0)\rangle\right] < \infty$.*

The first main result of this section is related to the tightness of the fluid-scaled state descriptor.

**Theorem 2.** *Suppose that Assumptions 1–3 are satisfied. Then, $\left\{(\overline{\boldsymbol{Q}}^{D,n}, \overline{\boldsymbol{Q}}^{S,n}, \overline{\boldsymbol{\eta}}^{D,n}, \overline{\boldsymbol{\eta}}^{S,n})\right\}_{n\in\mathbb{N}}$ is tight.*

Given the tightness result, the next two assumptions are used to prove that any subsequential limit is a fluid model solution $\mathbb{P}$-almost surely.

**Assumption 4.** *For each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, there exist $L_j^D < H_j^D$ and $L_k^S < H_k^S$ such that $h_j^D$ and $h_k^S$ are either bounded or lower-semicontinuous on $(L_j^D, H_j^D)$ and $(L_k^S, H_k^S)$, respectively.*

**Assumption 5.** *Assumptions 1-3 hold, and the following hold for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$,*

1. *$\overline{A}_j^D$ and $\overline{A}_k^S$ are continuous $\mathbb{P}$-almost surely, i.e., $\overline{A}_j^D$ and $\overline{A}_k^S$ are arrival functions $\mathbb{P}$-almost surely,*
2. *$\tilde{\eta}_j^D(0)$ and $\tilde{\eta}_k^S(0)$ do not charge points $\mathbb{P}$-almost surely, i.e., for any $x \in \mathbb{R}_+$, $\left\langle 1_{\{x\}}, \tilde{\eta}_j^D(0)\right\rangle = 0$ and $\left\langle 1_{\{x\}}, \tilde{\eta}_k^S(0)\right\rangle = 0$, $\mathbb{P}$-almost surely,*
3. *$\{\overline{M}_{jk}^n\}_{n\in\mathbb{N}}$ is $\mathcal{C}$-tight.*

**Theorem 3.** *If Assumptions 4 and 5 hold, then any distributional limit point $\{(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S)\}_{n\in\mathbb{N}}$ of $\{(\overline{\boldsymbol{Q}}^{D,n}, \overline{\boldsymbol{Q}}^{S,n}, \overline{\boldsymbol{\eta}}^{D,n}, \overline{\boldsymbol{\eta}}^{S,n})\}_{n\in\mathbb{N}}$ is $\mathbb{P}$-almost surely a fluid model solution for $(\overline{\boldsymbol{A}}^D, \overline{\boldsymbol{A}}^S)$ such that $(\overline{\boldsymbol{Q}}^D(0), \overline{\boldsymbol{Q}}^S(0), \overline{\boldsymbol{\eta}}^D(0), \overline{\boldsymbol{\eta}}^S(0))$ is equal in distribution to $(\tilde{\boldsymbol{Q}}^D(0), \tilde{\boldsymbol{Q}}^S(0), \tilde{\boldsymbol{\eta}}^D(0), \tilde{\boldsymbol{\eta}}^S(0))$ given in Assumption 3.*

In the remainder of this section, we present the proofs of Theorems 2 and 3. Section 4.1 identifies the compensator term that can be used to define martingales associated with the potential reneging marked point processes $\boldsymbol{\mathcal{S}}^{D,n}$ and $\boldsymbol{\mathcal{S}}^{S,n}$ for each $n \in \mathbb{N}$, and provides some preliminary results. Then, Section 4.2 contains the proof of Theorem 2, and Section 4.3 contains the proof of Theorem 3.

### 4.1 Preliminaries: Martingales and radon measures

The development in this section heavily leverages the presentation of the martingales associated with the potential reneging marked point processes in [43, Section 4.1] and is therefore kept concise, with proof details omitted.

Fix $n \in \mathbb{N}$. We start the analysis by defining a filtration $\{\mathcal{F}_t^n\}_{t\geqslant 0}$ such that for $t \geqslant 0$

$$\mathcal{F}_t^n = \sigma\left(\boldsymbol{Y}^n(0), \left(\boldsymbol{a}^{D,n}(s), 0 \leqslant s \leqslant t\right), \left(\boldsymbol{a}^{S,n}(s), 0 \leqslant s \leqslant t\right), \left(\boldsymbol{w}^{D,n}(s), 0 \leqslant s \leqslant t\right), \left(\boldsymbol{w}^{S,n}(s), 0 \leqslant s \leqslant t\right)\right),$$

where $\boldsymbol{a}^{D,n}$, $\boldsymbol{a}^{S,n}$, $\boldsymbol{w}^{D,n}$, and $\boldsymbol{w}^{S,n}$ are defined for the $n$th system as in Section 2. Note that $\mathcal{F}_t^n \subseteq \mathcal{G}_t^n$ for each $t \geqslant 0$, where $\mathcal{G}_t^n$ is given in Definition 2. Further, for each $j \in \mathbb{J}$, $k \in \mathbb{K}$, bounded measurable function $\phi : [0, H_j^D) \times \mathbb{R}_+ \to \mathbb{R}$, bounded measurable function $\psi : [0, H_k^S) \times \mathbb{R}_+ \to \mathbb{R}$, and $t \geqslant 0$, let

$$\mathcal{A}_j^{D,n}(\phi, t) = \int_0^t \left\langle \phi(\cdot, s) h_j^D(s), \eta_j^{D,n}(s) \right\rangle ds, \tag{41}$$

$$\mathcal{A}_k^{S,n}(\psi, t) = \int_0^t \left\langle \psi(\cdot, s) h_k^S(s), \eta_k^{S,n}(s) \right\rangle ds, \tag{42}$$

$$\mathcal{B}_j^{D,n}(\phi, t) = \mathcal{S}_j^{D,n}(\phi, t) - \mathcal{A}_j^{D,n}(\phi, t), \tag{43}$$

$$\mathcal{B}_k^{S,n}(\psi, t) = \mathcal{S}_k^{S,n}(\psi, t) - \mathcal{A}_k^{S,n}(\psi, t). \tag{44}$$

The following almost surely bounded and measurable functions will help us to bound the reneging processes. For $j \in \mathbb{J}$, $y \in [0, H_j^D)$, and $t \geqslant 0$, let

$$\theta_j^{D,n}(y, t) = 1_{[0, \chi_j^{D,n}(t-))}(y) \text{ and } \Theta_j^{D,n}(y, t) = 1_{[0, \chi_j^{D,n}(t-)]}(y). \tag{45}$$

For $k \in \mathbb{K}$, $y \in [0, H_k^S)$, and $t \geqslant 0$, let

$$\theta_k^{S,n}(y, t) = 1_{[0, \chi_k^{S,n}(t-))}(y) \text{ and } \Theta_k^{S,n}(y, t) = 1_{[0, \chi_k^{S,n}(t-)]}(y). \tag{46}$$

Definitions (45) and (46) and (23)–(26) lead to the following bounds for the reneging processes:

$$\mathcal{S}_j^{D,n}(\theta_j^{D,n}, t) \leqslant R_j^{D,n}(t) \leqslant \mathcal{S}_j^{D,n}(\Theta_j^{D,n}, t) \text{ and } \mathcal{S}_k^{S,n}(\theta_k^{S,n}, t) \leqslant R_k^{S,n}(t) \leqslant \mathcal{S}_k^{S,n}(\Theta_j^{S,n}, t). \tag{47}$$

Recall the independence of the initial conditions from the primitive processes and that the matching process $\boldsymbol{M}^n$ is $\{\mathcal{G}_t^n\}_{t\geqslant 0}$- adapted by Definition 2 for each $n \in \mathbb{N}$. By adapting the arguments used to prove [32, Corollary 5.5], Part 1 of [29, Proposition 5.1], and [29, Lemma 5.4] exactly as is mentioned in [43, Lemma 4], the following lemma holds.

**Lemma 3.** *Let $n \in \mathbb{N}$. For each $j \in \mathbb{J}$, $k \in \mathbb{K}$, bounded measurable function $\phi : [0, H_j^D) \times \mathbb{R}_+ \to \mathbb{R}$, such that $t \to \phi(w_{jl}^{D,n}(t), t)$ is left continuous on $[0, \infty)$ for each $l \in \left\{-\left\langle 1, \eta_j^{D,n}(0)\right\rangle + 1, \ldots, 0\right\} \cup \mathbb{N}$, and bounded measurable function $\psi : [0, H_k^S) \times \mathbb{R}_+ \to \mathbb{R}$ such that $t \to \psi(w_{kl}^{S,n}(t), t)$ is left continuous on $[0, \infty)$ for each $l \in \left\{-\left\langle 1, \eta_k^{S,n}(0)\right\rangle + 1, \ldots, 0\right\} \cup \mathbb{N}$, the processes $\mathcal{A}_j^{D,n}(\phi, \cdot)$ and $\mathcal{A}_k^{S,n}(\psi, \cdot)$ are the $\{\mathcal{F}_t\}_{t\geqslant 0}^n$-compensators of $\mathcal{S}_j^{D,n}(\phi, \cdot)$ and $\mathcal{S}_k^{S,n}(\psi, \cdot)$, respectively. Further, $\mathcal{A}_j^{D,n}(\Theta_j^{D,n}, \cdot)$ and $\mathcal{A}_k^{S,n}(\Theta_k^{S,n}, \cdot)$ are the $\{\mathcal{F}_t\}_{t\geqslant 0}^n$-compensators of $\mathcal{S}_j^{D,n}(\Theta_j^{D,n}, \cdot)$ and $\mathcal{S}_j^{S,n}(\Theta_j^{S,n}, \cdot)$. In particular, the processes $\mathcal{B}_j^{D,n}(\phi, \cdot)$, $\mathcal{B}_k^{S,n}(\psi, \cdot)$, $\mathcal{B}_j^{D,n}(\Theta_j^{D,n}, \cdot)$, and $\mathcal{B}_k^{S,n}(\Theta_k^{S,n}, \cdot)$ are local $\{\mathcal{F}_t^n\}_{t\geqslant 0}$-martingales.*

For $n \in \mathbb{N}$ a local $\{\mathcal{F}_t^n\}_{t\geqslant 0}$-martingale $L^n$, we denote by $\left\langle \overline{L}^n \right\rangle$ the quadratic variation process of the fluid-scaled process $\overline{L}^n = L^n/n$. The following result is the analogue of [43, Lemma 5].

**Lemma 4.** *Suppose that Assumptions 1 and 3 hold. For each $j \in \mathbb{J}$, $k \in \mathbb{K}$, $t \geqslant 0$, bounded measurable function $\phi : [0, H_j^D) \times \mathbb{R}_+ \to \mathbb{R}$ such that $u \mapsto \phi(w_{ji}^{D,n}(u), u)$ is left continuous on $[0,\infty)$ for each $i \in \left\{-\left\langle 1, \eta_j^{D,n}(0)\right\rangle + 1, \ldots, 0\right\} \cup \mathbb{N}$ and $n \in \mathbb{N}$, and bounded measurable function $\psi : [0, H_k^S) \times \mathbb{R}_+ \to \mathbb{R}$ such that $u \mapsto \psi(w_{ki}^{S,n}(u), u)$ is left continuous on $[0,\infty)$ for each $i \in \left\{-\left\langle 1, \eta_k^{S,n}(0)\right\rangle + 1, \ldots, 0\right\} \cup \mathbb{N}$ and $n \in \mathbb{N}$,*

$$\limsup_{n\to\infty} \mathbb{E}\left[\overline{H}^n(t)\right] < \infty,$$

*where $H^n(\cdot) = \mathcal{A}_j^{D,n}(\phi,\cdot)$, $\mathcal{A}_k^{S,n}(\psi,\cdot)$, $\mathcal{S}_j^{D,n}(\phi,\cdot)$, $\mathcal{S}_k^{S,n}(\psi,\cdot)$, $\mathcal{S}_j^{D,n}(\Theta_j^{D,n},\cdot)$, $\mathcal{S}_k^{S,n}(\Theta_k^{S,n},\cdot)$, $\mathcal{A}_j^{D,n}(\Theta_j^{D,n},\cdot)$, and $\mathcal{A}_k^{S,n}(\Theta_k^{S,n},\cdot)$. Moreover, for each $j \in \mathbb{J}$, $k \in \mathbb{K}$ and $t \geqslant 0$,*

$$\limsup_{n\to\infty} \mathbb{E}\left[\overline{R}_j^{D,n}(t)\right] < \infty \text{ and } \limsup_{n\to\infty} \mathbb{E}\left[\overline{R}_k^{S,n}(t)\right] < \infty.$$

*Furthermore, for $\overline{L}^n(\cdot) = \overline{\mathcal{B}}_j^{D,n}(\phi,\cdot)$, $\overline{\mathcal{B}}_k^{s,n}(\psi,\cdot)$, $\overline{\mathcal{B}}_j^{D,n}(\Theta_j^{D,n},\cdot)$, and $\overline{\mathcal{B}}_k^{S,n}(\Theta_k^{S,n},\cdot)$, for $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $n \in \mathbb{N}$, we have that for all $t \geqslant 0$*

$$\lim_{n\to\infty} \mathbb{E}\left[\left\langle \overline{L}^n \right\rangle(t)\right] = 0,$$

*and hence $\overline{L}^n \xrightarrow{d} 0$, as $n \to \infty$.*

We next provide alternative representations for the compensators of $\mathcal{S}_j^{D,n}(\theta_j^{D,n},\cdot)$, $\mathcal{S}_k^{S,n}(\theta_k^{S,n},\cdot)$, $\mathcal{S}_j^{D,n}(\Theta_j^{D,n},\cdot)$ and $\mathcal{S}_k^{S,n}(\Theta_k^{S,n},\cdot)$, for $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $n \in \mathbb{N}$. For $n \in \mathbb{N}$, $x \in \mathbb{R}_+$, and $t \geqslant 0$, define

$$F_{j,t}^{D,n}(x) = \left\langle 1_{[0,x]}, \eta_j^{D,n}(t)\right\rangle, j \in \mathbb{J}, \text{ and } F_{k,t}^{S,n}(x) = \left\langle 1_{[0,x]}, \eta_k^{S,n}(t)\right\rangle, k \in \mathbb{K}.$$

Also for $n \in \mathbb{N}$ and $t \geqslant 0$ define

$$\tilde{\chi}_j^{D,n}(t) := \inf\left\{x \in [0, H_j^D) : F_{j,t}^{D,n}(x) \geqslant \left\langle 1_{[0,\chi_j^{D,n}(t-))}, \eta_j^{D,n}(t)\right\rangle\right\}, j \in \mathbb{J},$$

and

$$\tilde{\chi}_k^{S,n}(t) := \inf\left\{x \in [0, H_k^{S,n}) : F_{k,t}^{S,n}(x) \geqslant \left\langle 1_{[0,\chi_k^{S,n}(t-))}, \eta_k^{S,n}(t)\right\rangle\right\}, k \in \mathbb{K}.$$

The following result is the analogue of [43, Lemma 6].

**Lemma 5.** *For each $n \in \mathbb{N}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, $t \geqslant 0$, $x \in [0, H_j^D)$, and $y \in [0, H_k^S)$, we have that*

$$\left\langle 1_{[0,x]} h_j^D, \eta_j^{D,n}(t)\right\rangle = \int_0^{F_{j,t}^{D,n}(x)} h_j^D\left(\left(F_{j,t}^{D,n}\right)^{-1}(s)\right) ds.$$

$$\left\langle 1_{[0,y]} h_k^S, \eta_k^{S,n}(t)\right\rangle = \int_0^{F_{k,t}^{S,n}(y)} h_k^S\left(\left(F_{k,t}^{S,n}\right)^{-1}(s)\right) ds.$$

*In particular, for each* $n \in \mathbb{N}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, *and* $t \geqslant 0$,

$$\mathcal{A}_j^{D,n}(\theta_j^{D,n}, t) = \int_0^t \int_0^{F_{j,t}^{D,n}(\tilde{\chi}_j^{D,n}(u))} h_j^D \left( \left( F_{j,u}^{D,n} \right)^{-1} (s) \right) ds du,$$

$$\mathcal{A}_j^{D,n}(\Theta_j^{D,n}, t) = \int_0^t \int_0^{F_{j,t}^{D,n}(\chi_j^{D,n}(u-))} h_j^D \left( \left( F_{j,u}^{D,n} \right)^{-1} (s) \right) ds du,$$

$$\mathcal{A}_k^{S,n}(\theta_j^{S,n}, t) = \int_0^t \int_0^{F_{k,t}^{S,n}(\tilde{\chi}_k^{S,n}(u))} h_k^S \left( \left( F_{k,u}^{S,n} \right)^{-1} (s) \right) ds du,$$

$$\mathcal{A}_k^{S,n}(\Theta_j^{S,n}, t) = \int_0^t \int_0^{F_{k,t}^{S,n}(\chi_k^{S,n}(u-))} h_k^S \left( \left( F_{k,u}^{S,n} \right)^{-1} (s) \right) ds du.$$

For each $n \in \mathbb{N}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $t \geqslant 0$, the measures $\mathcal{S}_j^{D,n}(\cdot, t)$ and $\mathcal{S}_k^{D,n}(\cdot, t)$ are finite Radon measures on $[0, H_j^D) \times \mathbb{R}_+$ and $[0, H_k^S) \times \mathbb{R}_+$, respectively. The next lemma shows that the corresponding compensators are bounded and its validity follows by the definition of the potential measures, (41), and (42), as in [43, Lemma 7].

**Lemma 6.** *For each* $n \in \mathbb{N}$, $j \in \mathbb{J}$, $k \in \mathbb{K}$, $0 < m < H_j^D$, $0 < u < H_k^S$, $t \geqslant 0$, *bounded measurable* $\phi : [0, H_j^s) \times \mathbb{R}_+ \to \mathbb{R}$, *such that* $\operatorname{supp}(\phi) \subseteq [0, m] \times \mathbb{R}_+$, *and bounded measurable* $\psi : [0, H_k^S) \times \mathbb{R}_+ \to \mathbb{R}$ *such that* $\operatorname{supp}(\psi) \subseteq [0, u] \times \mathbb{R}_+$, *we have that*

$$\left| \mathcal{A}_j^{D,n}(\phi, t) \right| \leqslant \|\phi\|_\infty \left( \left\langle 1, \eta_j^{D,n}(0) \right\rangle + A_j^{D,n}(t) \right) \int_0^m h_j^D(x) dx,$$

$$\left| \mathcal{A}_k^{S,n}(\psi, t) \right| \leqslant \|\psi\|_\infty \left( \left\langle 1, \eta_k^{S,n}(0) \right\rangle + A_k^{S,n}(t) \right) \int_0^u h_k^S(x) dx.$$

We remark that Lemmas 3, 5, and 6 all hold for fixed $n \in \mathbb{N}$, and, in particular, hold for the system model presented in Section 2. In contrast, Lemma 4 is an asymptotic result.

### 4.2 Proof of Theorem 2

Theorem 2 follows by Lemmas 7 and 8 below.

**Lemma 7.** *Suppose that Assumptions 1–3 hold. For any* $j \in \mathbb{J}$, $k \in \mathbb{K}$, $f_j \in \mathcal{C}_c^1([0, H_j^D))$, $h_k \in \mathcal{C}_c^1([0, H_k^S))$, $\phi_j \in \mathcal{C}_b([0, H_j^D) \times \mathbb{R}_+)$, $\psi_k \in \mathcal{C}_b([0, H_k^S) \times \mathbb{R}_+)$, *the sequences* $\{\overline{A}_j^{D,n}\}_{n\in\mathbb{N}}$, $\{\overline{A}_k^{S,n}\}_{n\in\mathbb{N}}$, $\{\overline{R}_j^{D,n}\}_{n\in\mathbb{N}}$, $\{\overline{R}_k^{S,n}\}_{n\in\mathbb{N}}$, $\{\overline{S}_j^{D,n}\}_{n\in\mathbb{N}}$, $\{\overline{S}_k^{S,n}\}_{n\in\mathbb{N}}$, $\{\overline{M}_{jk}^n\}_{n\in\mathbb{N}}$, $\left\{\left\langle 1, \overline{\eta}_j^{D,n}(\cdot) \right\rangle\right\}_{n\in\mathbb{N}}$, $\left\{\left\langle 1, \overline{\eta}_k^{S,n}(\cdot) \right\rangle\right\}_{n\in\mathbb{N}}$, $\left\{\left\langle f_j, \overline{\eta}_j^{D,n}(\cdot) \right\rangle\right\}_{n\in\mathbb{N}}$, $\left\{\left\langle h_k, \overline{\eta}_k^{S,n}(\cdot) \right\rangle\right\}_{n\in\mathbb{N}}$, $\{\overline{Q}_j^{D,n}\}_{n\in\mathbb{N}}$, $\{\overline{Q}_k^{S,n}\}_{n\in\mathbb{N}}$, $\left\{\overline{\mathcal{S}}_j^{D,n}(\phi_j, \cdot)\right\}_{n\in\mathbb{N}}$, $\left\{\overline{\mathcal{S}}_k^{S,n}(\psi_k, \cdot)\right\}_{n\in\mathbb{N}}$, $\left\{\overline{\mathcal{A}}_j^{D,n}(\phi_j, \cdot)\right\}_{n\in\mathbb{N}}$, *and* $\left\{\overline{\mathcal{A}}_k^{S,n}(\psi_k, \cdot)\right\}_{n\in\mathbb{N}}$, *are relatively compact in* $\mathcal{D}(\mathbb{R}_+)$, *and are therefore tight. If Assumption 5 also holds, then each of these processes is* $\mathcal{C}$*-tight.*

***Proof.*** Fix $j \in \mathbb{J}$, $k \in \mathbb{K}$, $f_j \in \mathcal{C}_c^1([0, H_j^D))$, $h_k \in \mathcal{C}_c^1([0, H_k^S))$, $\phi_j \in \mathcal{C}_b([0, H_j^D) \times \mathbb{R}_+)$, $\psi_k \in \mathcal{C}_b([0, H_k^S) \times \mathbb{R}_+)$.

The arrival processes $\left\{\overline{A}_j^{D,n}\right\}_{n\in\mathbb{N}}$ and $\left\{\overline{A}_j^{S,n}\right\}_{n\in\mathbb{N}}$ are relatively compact by the first condition of Assumption 1. Relative compactness of $\left\{\overline{S}_j^{D,n}\right\}_{n\in\mathbb{N}}$, $\left\{\overline{S}_k^{S,n}\right\}_{n\in\mathbb{N}}$, $\left\{\overline{\mathcal{A}}_j^{D,n}(\phi_j, \cdot)\right\}_{n\in\mathbb{N}}$, $\left\{\overline{\mathcal{A}}_k^{S,n}(\psi_k, \cdot)\right\}_{n\in\mathbb{N}}$,

$\left\{\overline{\mathcal{S}}_j^{D,n}(\phi_j,\cdot)\right\}_{n\in\mathbb{N}}$, $\left\{\overline{\mathcal{S}}_k^{S,n}(\psi_k,\cdot)\right\}_{n\in\mathbb{N}}$, $\left\{\left\langle 1,\overline{\eta}_j^{D,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$, and $\left\{\left\langle 1,\overline{\eta}_k^{S,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$ follows using Lemma 4 and the same arguments as in [29, Lemma 6.3]. For the reneging process $\left\{\overline{R}_j^{D,n}\right\}_{n\in\mathbb{N}}$, observe by (23) that $\overline{R}_j^{D,n}(t)\leqslant\overline{A}_j^{D,n}(t)$ for each $t\geqslant 0$ and $n\in\mathbb{N}$. Furthermore, for each $n\in\mathbb{N}$ and $0\leqslant s\leqslant t<\infty$, we have that

$$|\overline{R}_j^{D,n}(t)-\overline{R}_j^{D,n}(s)|\leqslant|\overline{S}_j^{D,n}(t)-\overline{S}_j^{D,n}(s)|.$$

Now, relative compactness of $\left\{\overline{R}_j^{D,n}\right\}_{n\in\mathbb{N}}$ follows by relative compactness of $\left\{\overline{S}_j^{D,n}\right\}_{n\in\mathbb{N}}$, and by exactly the same arguments $\left\{\overline{R}_k^{S,n}\right\}_{n\in\mathbb{N}}$ is relatively compact.

The matching process is relatively compact by Assumption 2, the fact that $\overline{M}_{jk}^n(t)\leqslant\min(\overline{Q}_j^{D,n}(0)+\overline{A}_j^{D,n}(t),\overline{Q}_k^{S,n}(0)+\overline{A}_k^{S,n}(t))$ for each $t\geqslant 0$ and $n\in\mathbb{N}$, and relative compactness of the arrival processes. By (21), observe that for $n\in\mathbb{N}$ and $0\leqslant s\leqslant t<\infty$,

$$|\overline{Q}_j^{D,n}(t)-\overline{Q}_j^{D,n}(s)|\leqslant|\overline{A}_j^{D,n}(t)-\overline{A}_j^{D,n}(s)|+|\overline{R}_j^{D,n}(t)-\overline{R}_j^{D,n}(s)|+\sum_{k\in\mathbb{K}}|\overline{M}_{jk}^n(t)-\overline{M}_{jk}^n(s)|.$$

Hence, $\left\{\overline{Q}_j^{D,n}\right\}_{n\in\mathbb{N}}$ is relatively compact by relative compactness of the arrival, reneging, and matching processes. Using (21), relative compactness of $\left\{\overline{Q}_k^{S,n}\right\}_{n\in\mathbb{N}}$ also follows.

Next, $\left\{\left\langle f_j,\overline{\eta}_j^{D,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$ and $\left\{\left\langle \phi_k,\overline{\eta}_k^{S,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$ are relatively compact by applying Lemma 1 and following arguments similar to those in [29, Lemma 6.4].

Finally, the sequences $\left\{\overline{\mathcal{A}}_j^{D,n}(\phi_j,\cdot)\right\}_{n\in\mathbb{N}}$ and $\left\{\overline{\mathcal{A}}_k^{S,n}(\psi_k,\cdot)\right\}_{n\in\mathbb{N}}$ are $\mathcal{C}$-tight because each process in the sequence is continuous. The sequences $\left\{\overline{R}_j^{D,n}\right\}_{n\in\mathbb{N}}$ and $\left\{\overline{R}_k^{S,n}\right\}_{n\in\mathbb{N}}$ are $\mathcal{C}$-tight because in the $n$th system each process in the sequence has jumps of size $1/n$ due to the continuity of the patience time distributions. If Assumption 5 holds, then parts 1 and 2 of Assumption 5 guarantee $\left\{\overline{A}_j^{D,n}\right\}_{n\in\mathbb{N}}$, $\left\{\overline{A}_k^{S,n}\right\}_{n\in\mathbb{N}}$, $\left\{\left\langle 1,\overline{\eta}_j^{D,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$, and $\left\{\left\langle 1,\overline{\eta}_k^{S,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$ are $\mathcal{C}$-tight. Part 3 of Assumption 5 guarantees $\left\{\overline{M}_{jk}^n\right\}_{n\in\mathbb{N}}$ is $\mathcal{C}$-tight. Then $\left\{\overline{Q}_j^{D,n}\right\}_{n\in\mathbb{N}}$ and $\left\{\overline{Q}_k^{S,n}\right\}_{n\in\mathbb{N}}$ are $\mathcal{C}$-tight from this and (21) and (22). $\mathcal{C}$-tightness of $\left\{\overline{\mathcal{S}}_j^{D,n}(\phi_j,\cdot)\right\}_{n\in\mathbb{N}}$ and $\left\{\overline{\mathcal{S}}_k^{S,n}(\psi_k,\cdot)\right\}_{n\in\mathbb{N}}$ follows when $\phi_j$ and $\psi_k$ are continuous, and so $\left\{\overline{S}_j^{D,n}\right\}_{n\in\mathbb{N}}$ and $\left\{\overline{S}_k^{S,n}\right\}_{n\in\mathbb{N}}$ are $\mathcal{C}$-tight (since the constant function 1 is continuous). Then, from (13) and (14) in Lemma 1, since $f_j$ and $h_k$ are continuous, $\left\{\left\langle f_j,\overline{\eta}_j^{D,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$, and $\left\{\left\langle h_k,\overline{\eta}_k^{S,n}(\cdot)\right\rangle\right\}_{n\in\mathbb{N}}$ are $\mathcal{C}$-tight. □

**Lemma 8.** *Suppose that Assumptions 1–3 hold. For each $j\in\mathbb{J}$ and $k\in\mathbb{K}$ the sequences $\{\overline{\eta}_j^{D,n}\}_{n\in\mathbb{N}}$, $\{\overline{\eta}_k^{S,n}\}_{n\in\mathbb{N}}$, $\{\overline{\mathcal{A}}_j^{D,n}\}_{n\in\mathbb{N}}$, $\{\overline{\mathcal{A}}_k^{S,n}\}_{n\in\mathbb{N}}$, $\{\overline{\mathcal{S}}_j^{D,n}\}_{n\in\mathbb{N}}$, and $\{\overline{\mathcal{S}}_k^{S,n}\}_{n\in\mathbb{N}}$, are relatively compact in $\mathcal{D}(\mathcal{M}[0,H_j^D))$, $\mathcal{D}(\mathcal{M}[0,H_k^S))$, $\mathcal{D}(\mathcal{M}([0,H_j^D)\times\mathbb{R}_+))$, $\mathcal{D}(\mathcal{M}([0,H_k^S)\times\mathbb{R}_+))$, $\mathcal{D}(\mathcal{M}([0,H_j^D)\times\mathbb{R}_+))$, and $\mathcal{D}(\mathcal{M}([0,H_k^S)\times\mathbb{R}_+))$, respectively.*

***Proof.*** Lemma 8 follows by Jabukbowski's criteria (see, e.g., [26, Theorem 4.6]) and using the same arguments as in [43, Lemma 9]. □

### 4.3 Proof of Theorem 3

For $n \in \mathbb{N}$, let

$$\overline{\boldsymbol{V}}^n = \left(\overline{\boldsymbol{A}}^{D,n}, \overline{\boldsymbol{A}}^{S,n}, \overline{\boldsymbol{Q}}^{D,n}, \overline{\boldsymbol{Q}}^{S,n}, \overline{\boldsymbol{\eta}}^{D,n}, \overline{\boldsymbol{\eta}}^{S,n}, \overline{\boldsymbol{R}}^{D,n}, \overline{\boldsymbol{R}}^{S,n}, \overline{\boldsymbol{M}}^n, \overline{\boldsymbol{\mathcal{S}}}^{D,n}, \overline{\boldsymbol{\mathcal{S}}}^{S,n}, \overline{\boldsymbol{\mathcal{A}}}^{D,n}, \overline{\boldsymbol{\mathcal{A}}}^{S,n}\right). \tag{48}$$

We use the following lemma to prove Theorem 3.

**Lemma 9.** *Suppose that Assumptions 4 and 5 hold, and that*

$$\overline{\boldsymbol{V}} = \left(\overline{\boldsymbol{A}}^{D}, \overline{\boldsymbol{A}}^{S}, \overline{\boldsymbol{Q}}^{D}, \overline{\boldsymbol{Q}}^{S}, \overline{\boldsymbol{\eta}}^{D}, \overline{\boldsymbol{\eta}}^{S}, \overline{\boldsymbol{R}}^{D}, \overline{\boldsymbol{R}}^{S}, \overline{\boldsymbol{M}}, \overline{\boldsymbol{\mathcal{S}}}^{D}, \overline{\boldsymbol{\mathcal{S}}}^{S}, \overline{\boldsymbol{\mathcal{A}}}^{D}, \overline{\boldsymbol{\mathcal{A}}}^{S}\right).$$

*is a distributional limit point of* $\left\{\overline{\boldsymbol{V}}^n\right\}_{n\in\mathbb{N}}$*, where* $\overline{\boldsymbol{V}}^n$ *is given in (48) for each* $n \in \mathbb{N}$*. Then, the following hold almost surely:*

1. *for each* $j \in \mathbb{J}$*,* $k \in \mathbb{K}$*,* $T > 0$*,* $u \in [0, H_j^D)$*, and* $m \in [0, H_k^S)$*, there exists* $L_j^D(u,T) < \infty$*,* $L_k^S(m,T) < \infty$ *such that for each* $\ell^D \in \boldsymbol{L}^1_{\mathrm{loc}}[0, H_j^D)$ *and* $\ell^S \in \boldsymbol{L}^1_{\mathrm{loc}}[0, H_k^S)$

$$\int_0^T \left\langle \ell^D, \eta_j^D(s) \right\rangle ds \leqslant L_j^D(u,T) \int_{[0,H_j^D)} \left|\ell^D(x)\right| dx,$$
$$\int_0^T \left\langle \ell^S, \eta_k^S(s) \right\rangle ds \leqslant L_k^S(m,T) \int_{[0,H_k^S)} \left|\ell^S(x)\right| dx;$$

2. *for all* $j \in \mathbb{J}$*,* $k \in \mathbb{K}$*,* $\phi \in \boldsymbol{C}_b([0, H_j^D) \times \mathbb{R}_+)$*,* $\psi \in \boldsymbol{C}_b([0, H_k^S) \times \mathbb{R}_+)$*, and* $t \geqslant 0$*,*

$$\mathcal{S}_j^D(\phi, t) = \mathcal{A}_j^D(\phi, t) = \int_0^t \left\langle \phi(\cdot, u) h_j^D(\cdot), \eta_j^D(u) \right\rangle du < \infty,$$
$$\mathcal{S}_k^S(\psi, t) = \mathcal{A}_k^S(\psi, t) = \int_0^t \left\langle \psi(\cdot, u) h_k^S(\cdot), \eta_k^S(u) \right\rangle du < \infty,$$

   *and, in particular, (31) holds;*

3. *for all* $j \in \mathbb{J}$*,* $k \in \mathbb{K}$*,* $t \geqslant 0$*, and* $x \in \mathbb{R}_+$*,* $\left\langle 1_{\{x\}}, \eta_j^D(t) \right\rangle = 0$ *and* $\left\langle 1_{\{x\}}, \eta_k^S(t) \right\rangle = 0$*, and, in particular, (32) holds;*

4. $\overline{\boldsymbol{R}}^D$ *and* $\overline{\boldsymbol{R}}^S$ *satisfy (33);*

5. $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S) \in \mathcal{C}(\overline{\mathbb{Y}})$*;*

6. $\overline{\boldsymbol{M}} \in \mathcal{C}(\mathbb{R}_+^{J \times K})$ *is a matching function;*

7. $\left(\overline{\boldsymbol{A}}^D, \overline{\boldsymbol{A}}^S, \overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S, \overline{\boldsymbol{R}}^D, \overline{\boldsymbol{R}}^S, \overline{\boldsymbol{M}}, \overline{\boldsymbol{\mathcal{A}}}^D, \overline{\boldsymbol{\mathcal{A}}}^S\right)$ *satisfy (34)-(35) and (36)-(37).*

***Proof.*** The proof of parts 1, 2, 3, and 4 follows analogously to the proof of [43, Lemma 10, parts 1,2,6,7] for the reneging measure $\eta$ in that paper (and ignoring the service measure $\nu$ in that paper). Part 5 follows from the following:

- The restriction (2) for each system $n$ in the sequence;

- The $\mathcal{C}$-tightness of the sequences $\{\overline{Q}_j^{D,n}\}_{n\in\mathbb{N}}, j \in \mathbb{J}$ and $\{\overline{Q}_k^{S,n}\}_{n\in\mathbb{N}}, k \in \mathbb{K}$ established in Lemma 7;
- The fact that $\overline{\eta}_j(t)$ and $\overline{\eta}_k(t)$ do not charge points for each $t \geqslant 0$ and $j \in \mathbb{J}$ and $k \in \mathbb{K}$ by part 3 above.

Part 6 follows from part 3 of Assumption 5 and the fact that $M_{jk}^n$ are non-decreasing for each $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $n \in \mathbb{N}$ from their definition in (18). To obtain part 7, we note the following:

- Arguments very similar to those used to establish [29, (3.11) for Theorem 7.1 (see page 51)] also show that the equations (34) and (35) hold, using Remark 1;
- The relations (36)-(37) follow by (21)-(22), and the convergence of the fluid-scaled processes.

□

***Proof of Theorem 3.*** Let $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S)$ be a distributional limit point of $\{(\overline{\boldsymbol{Q}}^{D,n}, \overline{\boldsymbol{Q}}^{S,n}, \overline{\boldsymbol{\eta}}^{D,n}, \overline{\boldsymbol{\eta}}^{S,n})\}_{n\in\mathbb{N}}$. Then there exists $\mathbb{N}' = \{n'\} \subseteq \mathbb{N}$ such that

$$(\overline{\boldsymbol{Q}}^{D,n'}, \overline{\boldsymbol{Q}}^{S,n'}, \overline{\boldsymbol{\eta}}^{D,n'}, \overline{\boldsymbol{\eta}}^{S,n'}) \Rightarrow (\overline{\boldsymbol{Q}}^{D}, \overline{\boldsymbol{Q}}^{S}, \overline{\boldsymbol{\eta}}^{D}, \overline{\boldsymbol{\eta}}^{S}), \text{ as } n' \to \infty.$$

For each $n \in \mathbb{N}$, let $\overline{V}^n$ be as given in (48) and consider the subsequence $\{\overline{V}^{n'}\}_{n'\in\mathbb{N}'}$. Let $\tilde{\boldsymbol{V}}$ be a limit point of $\{\overline{V}^{n'}\}_{n'\in\mathbb{N}'}$. Since $\{\overline{V}^{n'}\}_{n'\in\mathbb{N}'}$ is tight from Lemma 7 and Lemma 8, there exists a further subsequence $\mathbb{N}'' = \{n''\}$ such that $\overline{V}^{n''} \Rightarrow \tilde{V}$ as $n'' \to \infty$. Since $(\overline{\boldsymbol{Q}}^{D,n''}, \overline{\boldsymbol{Q}}^{S,n''}, \overline{\boldsymbol{\eta}}^{D,n''}, \overline{\boldsymbol{\eta}}^{S,n''})$ are coordinates of $\overline{V}^{n''}$ for each $n''$, it follows that $(\tilde{\boldsymbol{Q}}^D, \tilde{\boldsymbol{Q}}^S, \tilde{\boldsymbol{\eta}}^D, \tilde{\boldsymbol{\eta}}^S)$ is equal in distribution to $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S)$. Furthermore, by Lemma 9, $(\tilde{\boldsymbol{Q}}^D, \tilde{\boldsymbol{Q}}^S, \tilde{\boldsymbol{\eta}}^D, \tilde{\boldsymbol{\eta}}^S)$ is almost surely a fluid model solution. Hence the same is true for $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S)$. □

## 5 Stationarity

Here, we study the behavior of the stochastic model in stationarity. To ensure the existence of a stationary distribution, we make the following assumption.

**Assumption 6.** *The components of the arrival processes $\boldsymbol{A}^D$ and $\boldsymbol{A}^S$ are delayed renewal processes with absolutely continuous interarrival distributions that have finite means and the admissible matching policy $(\mathbb{X}, \{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\})$ is such that $\{\mathbb{P}_{\boldsymbol{y}} : \boldsymbol{y} \in \mathbb{X}\}$ is a time homogeneous Feller Markov process.*

As a consequence of Assumption 6, (1) holds. We also make the following assumption on the reneging distributions.

**Assumption 7.** *The reneging distributions satisfy the following conditions:*

1. $\int_0^\infty (1 - G_j^D(x))dx = 1/\theta_j^D \in (0, \infty)$, $j \in \mathbb{J}$, *and* $\int_0^\infty (1 - G_k^S(x))dx = 1/\theta_k^S \in (0, \infty)$, $k \in \mathbb{K}$;
2. $G_j^D$, $j \in \mathbb{J}$, *and* $G_k^S$, $k \in \mathbb{K}$ *are strictly increasing.*

Part 1 of Assumption 7 ensures that the reneging distributions have finite positive mean and Part 2 of Assumption 7 ensures that the inverse functions $\left(G_j^D\right)^{-1} : [0,1) \to [0, H_j^D)$ and $\left(G_k^S\right)^{-1} : [0,1) \to [0, H_k^S)$ are well-defined for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$. When Part 1 of Assumption 7 holds the excess life distributions of the reneging distributions are as follows:

$$G_{e,j}^D(x) = \int_0^x \theta_j^D(1 - G_j^D(u))du, \text{ for } j \in \mathbb{J} \text{ and } x \in \mathbb{R}_+$$

and

$$G_{e,k}^S(x) = \int_0^x \theta_k^S(1 - G_k^S(u))du, \text{ for } k \in \mathbb{K} \text{ and } x \in \mathbb{R}_+.$$

Under Assumptions 6 and 7, in Section 5.1, we show that the stochastic model admits at least one stationary distribution (see Proposition 1). Next, in Sections 5.2 and 5.3, we show that fixed points, also called invariant states, of the fluid model with suitable arrival functions are valid first order approximations for the stationary distributions of the stochastic system. Here suitable arrival functions are those with component functions that are absolutely continuous with constant densities. In Section 5.2, we restate two results from [6] that concern the invariant states of such a fluid model. The first (Proposition 2) gives a characterization of the invariant states. The second (Proposition 3) concerns the behavior of fluid model solutions as time becomes large and provides sufficient conditions for convergence to an invariant state. Then, in Section 5.3, we consider a sequence of stochastic systems in stationarity under fluid scaling. Under mild asymptotic conditions, we prove convergence of this sequence to an invariant state of the fluid model (see Theorem 4). The combination of these results establishes an interchange of limits result that justifies regarding invariant states as first order approximations to stationary distributions. This is illustrated in Figure 2 at the beginning of Section 5.3 after establishing the necessary notation.

## 5.1 Existence of a Stationary Distribution

In this section, we determine conditions under which the stochastic model admits a stationary distribution, although we make no claim about uniqueness of such a distribution.

**Proposition 1.** *Suppose that Assumption 6 and Part 1 of Assumption 7 hold. There exists $\xi \in \Xi$ such that the state process $\boldsymbol{Y}$ is stationary when $\boldsymbol{Y}(0)$ has distribution $\xi$. In particular, $\xi$ is such that for all $t \geqslant 0$*

$$\mathbb{E}_\xi\left[A_j^D(t)\right] = \lambda_j^D t \quad \text{and} \quad \mathbb{E}_\xi\left[\left\langle 1, \overline{\eta}_j^D(0)\right\rangle\right] = \lambda_j^D/\theta_j^D, \quad \text{for each } j \in \mathbb{J}, \tag{49}$$

$$\mathbb{E}_\xi\left[A_k^S(t)\right] = \lambda_k^S t \quad \text{and} \quad \mathbb{E}_\xi\left[\left\langle 1, \overline{\eta}_k^S(0)\right\rangle\right] = \lambda_k^S/\theta_k^S, \quad \text{for each } k \in \mathbb{K}. \tag{50}$$

***Proof.*** Fix $\boldsymbol{y} \in \mathbb{X}$. Let $\boldsymbol{Y}$ denote the state process with law $\mathcal{L}_{\boldsymbol{y}}$. For any Borel subset $B$ of $\mathbb{X}$, let $L_0(B) = \mathbb{P}_{\boldsymbol{y}}(\boldsymbol{Y}(0) \in B)$ and for $t > 0$ let

$$L_t(B) := \frac{1}{t}\int_0^t \mathbb{P}_{\boldsymbol{y}}(Y(s) \in B)ds.$$

Then $L_t$ is a Borel probability measure on $\mathbb{X}$ for each $t \geqslant 0$. From Assumption 6, the state process $\boldsymbol{Y}$ is a time homogeneous Feller Markov process. Also, since $\boldsymbol{y} \in \mathbb{X}$, $\max_{j\in\mathbb{J}}\left\langle 1, \eta_j^D(0)\right\rangle < \infty$ and $\max_{k\in\mathbb{K}}\left\langle 1, \eta_k^S(0)\right\rangle < \infty$ since $\boldsymbol{y} \in \mathbb{X} \subset \mathbb{Y}$. Then, upon recalling (1), we can argue very similarly

to the proof of [30, Lemma 4.8] (ignoring the service measure) to find that the family of measures $\{L_t\}_{t\geqslant 0}$ is tight. Finally, the Krylov–Bogoliubov theorem (see [17, Corollary 3.1.2]) implies that any limit point $\xi$ of $\{L_t\}_{t\geqslant 0}$ is a stationary distribution. The equations in (49) and (50) follow since the marginal distributions of the arrival processes and total mass of the potential queue measures must be stationary processes when the the initial condition is $\xi$ and the total mass of the potential queue measures is equal in distribution to that of an infinite server queue. □

### 5.2 Invariant States and Long Time Behavior of Fluid Model Solutions.

We restrict attention to fluid model solutions given in Definition 4 for which the arrival functions $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ are linear. Specifically, we suppose that for some $\boldsymbol{\lambda}^D \in (0,\infty)^J$ and $\boldsymbol{\lambda}^S \in (0,\infty)^K$ we have that for each $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $t \geqslant 0$,

$$\overline{A}^D_j(t) = \lambda^D_j t \qquad \text{and} \qquad \overline{A}^S_k(t) = \lambda^S_k t. \tag{51}$$

With this, we define invariant states as follows.

**Definition 5.** *Let $\boldsymbol{\lambda}^D \in (0,\infty)^J$ and $\boldsymbol{\lambda}^S \in (0,\infty)^K$. A tuple $(\boldsymbol{q}^{D,\star}, \boldsymbol{q}^{S,\star}, \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}) \in \overline{\mathbb{Y}}$ is an **invariant state** for $(\boldsymbol{\lambda}^D, \boldsymbol{\lambda}^S)$ if the constant function $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S)$ given by*

$$(\overline{\boldsymbol{Q}}^D(t), \overline{\boldsymbol{Q}}^S(t), \overline{\boldsymbol{\eta}}^D(t), \overline{\boldsymbol{\eta}}^S(t)) = (\boldsymbol{q}^{D,\star}, \boldsymbol{q}^{S,\star}, \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}), \text{ for all } t \geqslant 0 \tag{52}$$

*is a fluid model solution for $\left(\overline{\boldsymbol{A}}^D, \overline{\boldsymbol{A}}^S\right)$ given in (51). The **invariant manifold** for $(\boldsymbol{\lambda}^D, \boldsymbol{\lambda}^S)$ is the set of all invariant states for $(\boldsymbol{\lambda}^D, \boldsymbol{\lambda}^S)$, which we denote by $\mathcal{I}_{\boldsymbol{\lambda}}$.*

To build some intuition, fix $\boldsymbol{\lambda}^D \in (0,\infty)^J$ and $\boldsymbol{\lambda}^S \in (0,\infty)^K$ and suppose that an invariant state $(\boldsymbol{q}^{D,\star}, \boldsymbol{q}^{S,\star}, \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star})$ for $(\boldsymbol{\lambda}^D, \boldsymbol{\lambda}^S)$ exists. We begin by looking at the invariant potential queue measures since their evolution is independent of the matching function. By substituting the components of $\boldsymbol{\eta}^{D,\star}$ (resp. $\boldsymbol{\eta}^{S,\star}$) into dynamic equation (34) (resp. (35)) and letting $t$ tend to infinity, one can show that the term corresponding to the initial condition tends to zero and conclude that for each $j \in \mathbb{J}$ (resp. $k \in \mathbb{K}$), $\eta^{D,\star}_j$ (resp. $\eta^{S,\star}_k$) is absolutely continuous with density $\lambda^D_j(1 - G^D_j(\cdot))$ (resp. $\lambda^S_k(1 - G^S_k(\cdot))$).

Next, we look at the invariant fluid queue masses, which depend on the matching function $\overline{\boldsymbol{M}}$. By (33), for each $j \in \mathbb{J}$ (resp. $k \in \mathbb{K}$), $\overline{R}^D_j$ (resp. $\overline{R}^S_k$) is absolutely continuous with a constant density function. This together with (36) (resp. (37)) implies that for each $j \in \mathbb{J}$ (resp. $k \in \mathbb{K}$), $\overline{M}^D_j := \sum_{k\in\mathbb{K}} \overline{M}_{jk}$ (resp. $\overline{M}^S_k := \sum_{j\in\mathbb{J}} \overline{M}_{jk}$) is absolutely continuous with constant density function $\overline{m}^D_j \in [0, \lambda^d_j]$ (resp. $\overline{m}^S_k \in [0, \lambda^S_k]$). It follows that $\overline{M}_{jk}$, $j \in \mathbb{J}$ and $k \in \mathbb{K}$, are absolutely continuous, although their density functions aren't necessarily constant. However, it is straightforward to check that $\overline{\boldsymbol{M}}$ can be replaced with a matching function that has coordinate functions that are absolutely continuous with constant density. Let

$$\mathbb{M} := \left\{ \boldsymbol{m} \in \mathbb{R}^{J\times K}_+ : \sum_{k\in\mathbb{K}} m_{jk} \leqslant \lambda^D_j,\ j \in \mathbb{J} \text{ and } \sum_{j\in\mathbb{J}} m_{jk} \leqslant \lambda^S_k,\ k \in \mathbb{K}, \text{ and } m_{jk} = 0 \text{ for all } (j,k) \notin \mathcal{E} \right\}.$$

These observations implies that in order to characterize invariant states it suffices to restrict attention to matching functions $\overline{\boldsymbol{M}}$ such that $\overline{\boldsymbol{M}}(t) = \boldsymbol{m}t$ for all $t \geqslant 0$, for some $\boldsymbol{m} \in \mathbb{M}$.

The following results are proved in [6], and re-stated here for the reader's convenience.

**Proposition 2** (Proposition 3 in [6]). *Let $\boldsymbol{\lambda}^D \in (0,\infty)^J$ and $\boldsymbol{\lambda}^S \in (0,\infty)^K$ and suppose that Assumption 7 holds. A tuple $(\boldsymbol{q}^{D,\star}, \boldsymbol{q}^{S,\star}, \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}) \in \overline{\mathbb{Y}}$ is in the invariant manifold $\mathcal{I}_{\boldsymbol{\lambda}}$ if and only if it satisfies the following relations for some $\boldsymbol{m} \in \mathbb{M}$: For $j \in \mathbb{J}$, $k \in \mathbb{K}$, and $x \in \mathbb{R}_+$,*

$$\eta_j^{D,\star}(dx) = \lambda_j^D (1 - G_j^D(x))dx, \tag{53}$$

$$\eta_k^{S,\star}(dx) = \lambda_k^S (1 - G_k^S(x))dx, \tag{54}$$

$$q_j^{D,\star}(\boldsymbol{m}) = \begin{cases} \frac{\lambda_j^D}{\theta_j^D}, & \text{if } \sum_{k\in\mathbb{K}} m_{jk} = 0, \\ \frac{\lambda_j^D}{\theta_j^D} G_{e,j}^D\Big((G_j^D)^{-1}\Big(1 - \frac{\sum_{k\in\mathbb{K}} m_{jk}}{\lambda_j^D}\Big)\Big), & \text{if } \sum_{k\in\mathbb{K}} m_{jk} \in (0, \lambda_j^D], \end{cases} \tag{55}$$

$$q_k^{S,\star}(\boldsymbol{m}) = \begin{cases} \frac{\lambda_j^S}{\theta_j^S}, & \text{if } \sum_{j\in\mathbb{J}} m_{jk} = 0, \\ \frac{\lambda_j^S}{\theta_j^S} G_{e,k}^S\Big((G_k^S)^{-1}\Big(1 - \frac{\sum_{j\in\mathbb{J}} m_{jk}}{\lambda_k^S}\Big)\Big), & \text{if } \sum_{j\in\mathbb{J}} m_{jk} \in (0, \lambda_k^S]. \end{cases} \tag{56}$$

When a matching policy arising from some $\boldsymbol{m} \in \mathbb{M}$ is fixed, a fluid model solution approaches a unique invariant point, assuming "good" initial conditions.

**Proposition 3** (Theorem 2 in [6]). *For each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, assume $h_j^D$ and $h_k^S$ are bounded functions. Suppose that Assumption 7 is satisfied, $\overline{\boldsymbol{A}}^D$ and $\overline{\boldsymbol{A}}^S$ are arrival functions that satisfy (51), and $(\overline{\boldsymbol{Q}}^D, \overline{\boldsymbol{Q}}^S, \overline{\boldsymbol{\eta}}^D, \overline{\boldsymbol{\eta}}^S) \in \mathcal{C}(\overline{\mathbb{Y}})$ is a fluid model solution for $(\overline{\boldsymbol{A}}^D, \overline{\boldsymbol{A}}^S)$ such that $\overline{\boldsymbol{\eta}}_j^D(0)$, $j \in \mathbb{J}$, and $\overline{\boldsymbol{\eta}}_k^S(0)$, $k \in \mathbb{K}$, do not charge points and the matching function $\overline{\boldsymbol{M}}$ is such that $\overline{\boldsymbol{M}}(t) = \boldsymbol{m}t$, $t \geqslant 0$ for some $\boldsymbol{m} \in \mathbb{M}$. Then*

$$\lim_{t\to\infty} \left(\overline{\boldsymbol{Q}}^D(t), \overline{\boldsymbol{Q}}^S(t), \overline{\boldsymbol{\eta}}^D(t), \overline{\boldsymbol{\eta}}^S(t)\right) = \left(\boldsymbol{q}^{D,\star}(\boldsymbol{m}), \boldsymbol{q}^{S,\star}(\boldsymbol{m}), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right).$$

**Remark 3.** *The requirement in Proposition 3 that the matching function has constant matching rate is restrictive, and is of interest to relax. However, that is beyond the scope of the present paper.*

### 5.3 Convergence of fluid scaled stationary states to the invariant manifold

In this section, we consider a sequence of stationary stochastic systems indexed by $n \in \mathbb{N}$.

**Assumption 8.** *For each $n \in \mathbb{N}$, there is a stochastic system indexed by $n$ such that the arrival processes $\boldsymbol{A}^{D,n}$ and $\boldsymbol{A}^{S,n}$ and admissible matching policy $(\mathbb{X}^n, \{\mathbb{P}_{\boldsymbol{y}}^n : \boldsymbol{y} \in \mathbb{X}^n\})$ satisfy Assumption 6, while the stochastic primitives do not depend on $n$ and the reneging distributions satisfy Assumption 7.*

If Assumption 8 holds, then, by Proposition 1, a stationary distribution $\varsigma^n \in \Xi^n$ exists for each $n \in \mathbb{N}$ and we let $\boldsymbol{Y}^n(\infty) = (\boldsymbol{a}^{D,n}(\infty), \boldsymbol{a}^{S,n}(\infty), \boldsymbol{Q}^{D,n}(\infty), \boldsymbol{Q}^{S,n}(\infty), \boldsymbol{\eta}^{D,n}(\infty), \boldsymbol{\eta}^{S,n}(\infty))$ denote a stationary state for the $n^{th}$ system that has distribution $\varsigma^n$. We apply fluid scaling to this sequence of stochastic stationary states and provide conditions under which the fluid scaled sequence of stationary states converges to an invariant state, as shown in Theorem 4 below. As a consequence under suitable conditions that in particular imply ergodicity for the stochastic systems, the fluid ($n \to \infty$) and stationary ($t \to \infty$) limits can be interchanged as illustrated in Figure 2.

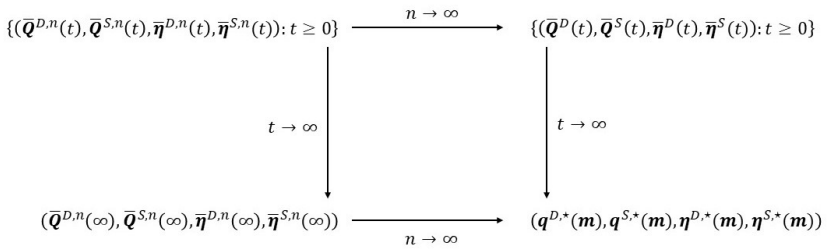


Figure 2: When Assumption 9 holds, the arrival functions $\overline{\mathbf{A}}^D$ and $\overline{\mathbf{A}}^S$ for the fluid model are such that $\overline{A}^D_j(t) = \lambda^S_j t$, $j \in \mathbb{J}$, and $\overline{A}^S_k(t) = \lambda^S_k t$, $k \in \mathbb{K}$, for all $t \geqslant 0$, and the sequence of matching functions is well behaved (see condition in Theorem 4), Propositions 1 and 3 and Theorems 3 and 4, ensure that when the limits exists, the limits $t \to \infty$ and $n \to \infty$ can be taken in either order, as illustrated in the figure, where $\mathbf{m} \in \mathbb{M}$.

In order to prove Theorem 4 below, we consider a sequence of systems indexed by $n \in \mathbb{N}$ as in Section 4 such that the initial condition $\boldsymbol{Y}^n(0)$ for the $n^{th}$ system has distribution $\varsigma^n$, i.e., has a stationary distribution. Henceforth for each $n \in \mathbb{N}$, $\boldsymbol{Y}^n(\cdot) = \left(\boldsymbol{a}^{D,n}(\cdot), \boldsymbol{a}^{S,n}(\cdot), \boldsymbol{Q}^{D,n}(\cdot), \boldsymbol{Q}^{S,n}(\cdot), \boldsymbol{\eta}^{D,n}(\cdot), \boldsymbol{\eta}^{S,n}(\cdot)\right)$ denotes the state process with initial condition $\boldsymbol{Y}^n(0)$ that has distribution $\varsigma^n$, i.e., $\boldsymbol{Y}^n(0)$ has a stationary distribution so that $\boldsymbol{Y}^n$ is a stationary process. We aim to apply Theorem 3 to the sequence of $\left\{\left(\overline{\boldsymbol{Q}}^{D,n}, \overline{\boldsymbol{Q}}^{S,n}, \overline{\boldsymbol{\eta}}^{D,n}, \overline{\boldsymbol{\eta}}^{S,n}\right)\right\}_{n\in\mathbb{N}}$ of fluid scaled stationary processes. To this end, we require the sequence of arrival processes to satisfy the following assumption.

**Assumption 9.** *Suppose that Assumption 8 holds and for some $\boldsymbol{\lambda}^D \in (0,\infty)^J$ and $\boldsymbol{\lambda}^S \in (0,\infty)^K$ we have*

$$\lim_{n\to\infty} \frac{\boldsymbol{\lambda}^{D,n}}{n} = \boldsymbol{\lambda}^D \qquad \text{and} \qquad \lim_{n\to\infty} \frac{\boldsymbol{\lambda}^{S,n}}{n} = \boldsymbol{\lambda}^S,$$

*where for each $n \in \mathbb{N}$, $1/\lambda^{D,n}_j$ denotes the mean interarrival time of $A^{D,n}_j$, $j \in \mathbb{J}$, and $1/\lambda^{S,n}_k$ denotes the mean interarrival time of $A^{D,n}_k$, $k \in \mathbb{K}$.*

When Assumption 9 holds, it follows that Assumption 1 and Part 1 of Assumption 5 hold for the arrival functions $\overline{\mathbf{A}}^D$ and $\overline{\mathbf{A}}^S$ such that $\overline{A}^D_j(t) = \lambda^D_j t$, $j \in \mathbb{J}$, and $\overline{A}^S_k(t) = \lambda^S_k t$, $k \in \mathbb{K}$, for all $t \geqslant 0$.

In preparation for proving Theorem 4, we establish a tightness result. Our proofs leverage results for a many-server queue with reneging in [30], which were also leveraged in [49], to prove convergence of stationary distributions. The key observation is that the potential queue measure in [30] does not depend on the service process, and so results on that measure can be carried over to this paper since the potential queue measures here do not depend on the matching process. As in Section 4 for each $n \in \mathbb{N}$, we will use $\mathbb{P}$ and $\mathbb{E}$ in place of $\mathbb{P}^n_{\varsigma^n}$ and $\mathbb{E}^n_{\varsigma^n}$ respectively to simplify the notation.

**Lemma 10.** *Suppose Assumption 9 holds (which implies that Assumptions 6, 7 and 8 hold). Then, $\left\{\left(\overline{\boldsymbol{Q}}^{D,n}(\infty), \overline{\boldsymbol{Q}}^{S,n}(\infty), \overline{\boldsymbol{\eta}}^{D,n}(\infty), \overline{\boldsymbol{\eta}}^{S,n}(\infty)\right)\right\}_{n\in\mathbb{N}}$ is tight. In addition, as $n \to \infty$,*

$$\left(\overline{\boldsymbol{\eta}}^{D,n}(\infty), \overline{\boldsymbol{\eta}}^{S,n}(\infty)\right) \xrightarrow{d} \left(\boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right), \tag{57}$$

*and, for each $j \in \mathbb{J}$ and $k \in \mathbb{K}$,*

$$\lim_{n\to\infty} \mathbb{E}\left[\langle 1, \overline{\eta}_j^{D,n}(\infty)\rangle\right] = \frac{\lambda_j^D}{\theta_j^D} = \langle 1, \eta_j^{D,\star}\rangle \quad \textit{and} \quad \lim_{n\to\infty} \mathbb{E}\left[\langle 1, \overline{\eta}_k^{S,n}(\infty)\rangle\right] = \frac{\lambda_k^S}{\theta_k^S} = \langle 1, \eta_k^{S,\star}\rangle. \tag{58}$$

*Finally, if $\{n_l\}_{l\in\mathbb{N}}$ is a strictly increasing subsequence of $\mathbb{N}$ and*

$$\left(\overline{\boldsymbol{Q}}^{D,n_l}(\infty), \overline{\boldsymbol{Q}}^{S,n_l}(\infty), \overline{\boldsymbol{\eta}}^{D,n_l}(\infty), \overline{\boldsymbol{\eta}}^{S,n_l}(\infty)\right) \xrightarrow{d} \left(\overline{\boldsymbol{Q}}^{D}(\infty), \overline{\boldsymbol{Q}}^{S}(\infty), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right), \qquad \textit{as } l \to \infty,$$

*then*

$$\lim_{l\to\infty} \mathbb{E}\left[\overline{Q}_j^{D,n_l}(\infty)\right] = \mathbb{E}\left[Q_j^D(\infty)\right] \quad \textit{and} \quad \lim_{l\to\infty} \mathbb{E}\left[\overline{Q}_k^{S,n_l}(\infty)\right] = \mathbb{E}\left[\overline{Q}_k^{S}(\infty)\right]. \tag{59}$$

***Proof.*** The same arguments used to prove the tightness of the scaled potential queue measure in [30, Theorem 6.2] establish that the families $\left\{\overline{\eta}_j^{D,n}(\infty)\right\}_{n\in\mathbb{N}}$ and $\left\{\overline{\eta}_k^{S,n}(\infty)\right\}_{n\in\mathbb{N}}$ are tight since the potential queue measures are independent of the matching policy. The argument to see that (57) holds is identical to the argument in the proof of [30, Theorem 3.3, the paragraph surrounding (6.15)] that shows the measure $\tilde{\eta}_*$ satisfies (6.15). Then (58) follows from (57), (49), (50) and the definitions of $\boldsymbol{\eta}^{D,\star}$ and $\boldsymbol{\eta}^{S,\star}$.

The tightness of $\left\{\overline{Q}_j^{D,n}(\infty)\right\}_{n\in\mathbb{N}}$, $j \in \mathbb{J}$, follows by first observing that (2) implies

$$\mathbb{E}\left[\overline{Q}_j^{D,n}(\infty)\right] \leqslant \mathbb{E}\left[\left\langle 1, \overline{\eta}_j^{D,n}(\infty)\right\rangle\right].$$

Hence, for each $j \in \mathbb{J}$,

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left[\overline{Q}_j^{D,n}(\infty)\right] \leqslant \sup_{n\in\mathbb{N}} \frac{\lambda_j^{D,n}}{n\theta_j} < \infty,$$

by recalling from Assumption 9 that $\lim_{n\to\infty} \lambda_j^{D,n}/n = \lambda_j^D$. Further, by Markov's inequality, for each $j \in \mathbb{J}$, we find that

$$\lim_{c\to\infty} \mathbb{P}\left(\overline{Q}_j^{D,n}(\infty) > c\right) \leqslant \lim_{c\to\infty} \frac{\mathbb{E}\left[\overline{Q}_j^{D,n}(\infty)\right]}{c} \leqslant \lim_{c\to\infty} \frac{\lambda_j^{D,n}}{n}\frac{1}{c} = 0,$$

which shows the desired result. The tightness of $\left\{\overline{Q}_k^{S,n}(\infty)\right\}_{n\in\mathbb{N}}$, $k \in \mathbb{K}$, follows by the exact same argument. Tightness of $\left\{\left(\overline{\boldsymbol{Q}}^{D,n}(\infty), \overline{\boldsymbol{Q}}^{S,n}(\infty), \overline{\boldsymbol{\eta}}^{D,n}(\infty), \overline{\boldsymbol{\eta}}^{S,n}(\infty)\right)\right\}_{n\in\mathbb{N}}$ now follows. Finally, (59) follows from the dominated convergence theorem. □

In what follows, $\stackrel{d}{=}$ denotes equality in distribution and $\xrightarrow{d}$ denotes convergence in distribution. Also, for each $n \in \mathbb{N}$, let $\boldsymbol{M}^n$ denote the matching process given by (21) and (22) for the stationary process $\boldsymbol{Y}^n$ with initial condition $\boldsymbol{Y}^n(0)$ that has stationary initial distribution $\varsigma^n$.

**Theorem 4.** *For each $j \in \mathbb{J}$ and $k \in \mathbb{K}$, assume $h_j^D$ and $h_k^S$ are bounded functions. Suppose that Assumptions 4 and 9 hold (which implies that Assumptions 6, 7 and 8 hold) and $\mathbb{P}$-almost surely*

$$\lim_{n\to\infty} \overline{M}_{jk}^n = \overline{M}_{jk} \textit{ for each } j \in \mathbb{J} \textit{ and } k \in \mathbb{K}, \tag{60}$$

*where for some $\boldsymbol{m} \in \mathbb{M}$, $\overline{M}_{jk}(t) = m_{jk}t$ for all $t \geqslant 0$, $j \in \mathbb{J}$ and $k \in \mathbb{K}$. Then, as $n \to \infty$,*

$$\left(\overline{\boldsymbol{Q}}^{D,n}(\infty), \overline{\boldsymbol{Q}}^{S,n}(\infty), \overline{\boldsymbol{\eta}}^{D,n}(\infty), \overline{\boldsymbol{\eta}}^{S,n}(\infty)\right) \xrightarrow{d} \left(\boldsymbol{q}^{D,\star}(\boldsymbol{m}), \boldsymbol{q}^{S,\star}(\boldsymbol{m}), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right) \in \mathcal{I}_{\boldsymbol{\lambda}}.$$

***Proof.*** We begin by verifying that the conditions in Theorem 3 hold along subsequences of $\left\{\left(\overline{\boldsymbol{Q}}^{D,n}, \overline{\boldsymbol{Q}}^{S,n}, \overline{\boldsymbol{\eta}}^{D,n}, \overline{\boldsymbol{\eta}}^{S,n}\right)\right\}_{n\in\mathbb{N}}$ such that the sequence of initial conditions is convergent. For this, recall that Assumption 9 is sufficient to ensure that Assumption 1 and Part 1 of Assumption 5 are satisfied. In addition, by (60), Part 3 of Assumption 5 holds, which implies that Assumption 2 holds as well. Also, for each $n \in \mathbb{N}$,

$$\left(\overline{\boldsymbol{Q}}^{D,n}(0), \overline{\boldsymbol{Q}}^{S,n}(0), \overline{\boldsymbol{\eta}}^{D,n}(0), \overline{\boldsymbol{\eta}}^{S,n}(0)\right) \stackrel{d}{=} \left(\overline{\boldsymbol{Q}}^{D,n}(\infty), \overline{\boldsymbol{Q}}^{S,n}(\infty), \overline{\boldsymbol{\eta}}^{D,n}(\infty), \overline{\boldsymbol{\eta}}^{S,n}(\infty)\right). \tag{61}$$

Hence, by (61) and Lemma 10, there exists a strictly increasing subsequence $\{n_l\}_{l\in\mathbb{N}} \subseteq \mathbb{N}$ and $\left(\overline{\boldsymbol{Q}}^{D}(\infty), \overline{\boldsymbol{Q}}^{S}(\infty), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right) \in \overline{\mathbb{Y}}$ such that

$$\left(\overline{\boldsymbol{Q}}^{D,n_l}(0), \overline{\boldsymbol{Q}}^{S,n_l}(0), \overline{\boldsymbol{\eta}}^{D,n_l}(0), \overline{\boldsymbol{\eta}}^{S,n_l}(0)\right) \stackrel{d}{\to} \left(\overline{\boldsymbol{Q}}^{D}(\infty), \overline{\boldsymbol{Q}}^{S}(\infty), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right), \qquad \text{as } l \to \infty. \tag{62}$$

Without loss of generality, we may assume that this convergence is $\mathbb{P}$-almost sure. Then, also by (61) and Lemma 10, $\left\{\left(\overline{\boldsymbol{Q}}^{D,n_l}(0), \overline{\boldsymbol{Q}}^{S,n_l}(0), \overline{\boldsymbol{\eta}}^{D,n_l}(0), \overline{\boldsymbol{\eta}}^{S,n_l}(0)\right)\right\}_{l\in\mathbb{N}}$ satisfies Assumption 3 and $\left(\boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right)$ satisfies Part 2 of Assumption 5. In summary, Assumption 5 holds. Thus, when considering that Assumption 4 holds by the statement of Theorem 4, Theorem 3 implies that

$$(\overline{\boldsymbol{Q}}^{D,n_l}, \overline{\boldsymbol{Q}}^{S,n_l}, \overline{\boldsymbol{\eta}}^{D,n_l}, \overline{\boldsymbol{\eta}}^{S,n_l}) \stackrel{d}{\to} (\overline{\boldsymbol{Q}}^{D}, \overline{\boldsymbol{Q}}^{S}, \overline{\boldsymbol{\eta}}^{D}, \overline{\boldsymbol{\eta}}^{S}), \text{ as } l \to \infty,$$

where the limit point $\left(\overline{\boldsymbol{Q}}^{D}, \overline{\boldsymbol{Q}}^{S}, \overline{\boldsymbol{\eta}}^{D}, \overline{\boldsymbol{\eta}}^{S}\right)$ is $\mathbb{P}$-almost surely a fluid model solution for $(\overline{\boldsymbol{A}}^{D}, \overline{\boldsymbol{A}}^{S})$ such that

$$(\overline{\boldsymbol{Q}}^{D}(0), \overline{\boldsymbol{Q}}^{S}(0), \overline{\boldsymbol{\eta}}^{D}(0), \overline{\boldsymbol{\eta}}^{S}(0)) \stackrel{d}{=} \left(\overline{\boldsymbol{Q}}^{D}(\infty), \overline{\boldsymbol{Q}}^{S}(\infty), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right). \tag{63}$$

Moreover, due to (60), the matching function $\boldsymbol{M}$ satisfies $\overline{\boldsymbol{M}}(t) = \boldsymbol{m}t$ with $\boldsymbol{m} \in \mathbb{M}$ for each $t \geqslant 0$, $\mathbb{P}$-almost surely. Then, by Proposition 3, $\mathbb{P}$-almost surely,

$$\lim_{t\to\infty} (\overline{\boldsymbol{Q}}^{D}(t), \overline{\boldsymbol{Q}}^{S}(t), \overline{\boldsymbol{\eta}}^{D}(t), \overline{\boldsymbol{\eta}}^{S}(t)) = (\boldsymbol{q}^{D,\star}(\boldsymbol{m}), \boldsymbol{q}^{S,\star}(\boldsymbol{m}), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}). \tag{64}$$

However, by stationarity, for each $t \geqslant 0$,

$$(\overline{\boldsymbol{Q}}^{D}(t), \overline{\boldsymbol{Q}}^{S}(t), \overline{\boldsymbol{\eta}}^{D}(t), \overline{\boldsymbol{\eta}}^{S}(t)) \stackrel{d}{=} \left(\overline{\boldsymbol{Q}}^{D}(\infty), \overline{\boldsymbol{Q}}^{S}(\infty), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right). \tag{65}$$

Thus, combining (64) and (65) we see that

$$\left(\overline{\boldsymbol{Q}}^{D}(\infty), \overline{\boldsymbol{Q}}^{S}(\infty), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}\right) \stackrel{d}{=} (\boldsymbol{q}^{D,\star}(\boldsymbol{m}), \boldsymbol{q}^{S,\star}(\boldsymbol{m}), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}).$$

Combining this with (61) and (62), we see that

$$\left(\overline{\boldsymbol{Q}}^{D,n_l}(\infty), \overline{\boldsymbol{Q}}^{S,n_l}(\infty), \overline{\boldsymbol{\eta}}^{D,n_l}(\infty), \overline{\boldsymbol{\eta}}^{S,n_l}(\infty)\right) \stackrel{d}{\to} (\boldsymbol{q}^{D,\star}(\boldsymbol{m}), \boldsymbol{q}^{S,\star}(\boldsymbol{m}), \boldsymbol{\eta}^{D,\star}, \boldsymbol{\eta}^{S,\star}), \qquad \text{as } l \to \infty.$$

Since $\left\{\left(\overline{\boldsymbol{Q}}^{D,n}(\infty), \overline{\boldsymbol{Q}}^{S,n}(\infty), \overline{\boldsymbol{\eta}}^{D,n}(\infty), \overline{\boldsymbol{\eta}}^{S,n}(\infty)\right)\right\}_{n\in\mathbb{N}}$ is tight, the proof is complete. □

## Acknowledgments

We would like to thank Levi DeValve for helpful discussion related to the use of the fluid model invariant states to formulate and solve a matching policy optimization problem, which resulted in the paper [6]. Financial support from the University of Chicago Booth School of Business is gratefully acknowledged.

## References

[1] P. Afèche, A. Diamont, and J. Milner. Double-sided batch queues with abandonments: Modeling crossing networks. *Probability in the Engineering and Informational Sciences*, 25(2):135–155, 2011.

[2] P. Agarwal and K. Ramanan. Invariant states of hydrodynamic limits of randomized load balancing networks, 2020. arXiv:2008.08510.

[3] R. Aghajani and K. Ramanan. The hydrodynamic limit of a randomized load balancing network. *Ann. Appl. Probab.*, 29(4):2114–2174, 2019.

[4] N. Arnosti, R. Johari, and Y. Kanoria. Managing congestion in matching markets. *Manufacturing & Service Operations Management*, 23(3):620–636, 2021.

[5] R. Atar, H. Kaspi, and N. Shimkin. Fluid limits for many-server systems with reneging under a priority policy. *Math. Oper. Res.*, 39(3):672–696, 2014.

[6] A. Aveklouris, L. DeValve, and A. R. Ward. Matching impatient and heterogeneous demand and supply. *ArXiv preprint ArXiv:2102.02710*, 2021.

[7] S. Banerjee, A. Budhiraja, and A. L. Puha. Heavy traffic scaling limits for shortest remaining processing time queues with heavy tailed processing time distributions. *Ann. Appl. Probab.*, 32(4):2587–2651, 2022.

[8] S. Banerjee, Y. Kanoria, and P. Qian. State dependent control of closed queueing networks with application to ride-hailing, 2018. ArXiv preprint arXiv:1803.04959.

[9] S. Benjaafar and M. Hu. Operations management in the age of the sharing economy: What is old and what is new? *Manufacturing & Service Operations Management*, 22(1):93–101, 2020.

[10] P. Billingsley. *Probability and Measure. Wiley Series in Probability and Mathematical Statistics.* Wiley, New York, third edition, 1995.

[11] P. Billingsley. *Convergence of probability measures.* Wiley, New York, second edition, 1999.

[12] J. H. Blanchet, M. I. Reiman, V. Shah, L. M. Wein, and L. Wu. Asymptotically optimal control of a centralized dynamic matching market with general utilities. *Operations Research*, 70(6):3355–3370, 2022.

[13] O. J. Boxma, I. David, D. Perry, and W. Stadje. A new look at organ transplantation models and double matching queues. *Probability in the Engineering and Informational Sciences*, 25(2):135–155, 2011.

[14] B. Büke and H. Chen. Fluid and diffusion approximations of probabilistic matching systems. *Queueing Systems*, 86(1-2):1–33, 2017.

[15] F. Castro, H. Nazerzadeh, and C. Yan. Matching queues with reneging: A product form solution. *Queueing Systems*, 96(3-4):359–385, 2020.

[16] Y.-J. Chen, T. Dai, C. G. Korpeoglu, E. Körpeoğlu, O. Sahin, C. S. Tang, and S. Xiao. OM forum—innovative online platforms: Research opportunities. *Manufacturing & Service Operations Management*, 22(3):430–445, 2020.

[17] G. Da Prato, J. Zabczyk, and J. Zabczyk. *Ergodicity for infinite dimensional systems*, volume 229. Cambridge University Press, 1996.

[18] Y. Ding, S. T. McCormick, and M. Nagarajan. A fluid model for one-sided bipartite matching queues with match-dependent rewards. *Operations Research*, 69(4):1256–1281, 2021.

[19] D. Down, H. C. Gromoll, and A. L. Puha. Fluid limits for shortest remaining processing time queues. *Mathematics of Operations Research*, 34:880–911, 2009.

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

[21] C. Gromoll, P. Robert, and B. Zwart. Fluid limits for processor-sharing queues with impatience. *Mathematics of Operations Research*, 33(2):375–402, 2008.

[22] C. Gromoll and R. Williams. Fluid limits for networks with bandwidth sharing and general document size distributions. *The Annals of Applied Probability*, 19(1):243–280, 2009.

[23] H. C. Gromoll, L. Kruk, and A. L. Puha. The diffusion limit of an SRPT queue. *Stochsatic Systems*, 1:1–16, 2011.

[24] M. Hu, editor. *Sharing economy: making supply meet demand.* Springer Series in Supply Chain Management, 2019.

[25] M. Hu. From the classics to new tunes: A neoclassical view on sharing economy and innovative marketplaces. *Production and Operations Management*, 30(6):1668–1685, 2021.

[26] A. Jakubowski. On the Skorokhod topology. *Ann. Inst. H. Poincare Probab. Statist*, 22(3):263–285, 1986.

[27] M. Jonckheere, P. Moyal, C. Ramírez, and N. Soprano-Loto. Generalized max-weight policies in stochastic matching. *Stochastic Systems*, 13(1):40–58, 2023.

[28] W. Kang. Fluid limits of many-server retrial queues with nonpersistent customers. *Queueing Systems*, 79(2):183–219, 2015.

[29] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *The Annals of Applied Probability*, 20(6):2204–2260, 2010.

[30] W. Kang and K. Ramanan. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *The Annals of Applied Probability*, 22(2):477–521, 2012.

[31] Y. Kanoria and D. Saban. Facilitating the search for partners on matching platforms. *Management Science*, 67(10):5990–6029, 2021.

[32] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *The Annals of Applied Probability*, 21(1):33–114, 2011.

[33] A. Khademi and X. Liu. Asymptotically optimal allocation policies for transplant queueing systems. *SIAM Journal on Applied Mathematics*, 81(3):1116–1140, 2021.

[34] A. Kohlenberg and I. Gurvich. The cost of impatience in dynamic matching: Scaling laws and operating regimes, 2023. Available at SSRN 4453900.

[35] V. Limic. A LIFO queue in heavy traffic. *Ann. Appl. Probab.*, 11(2):301–331, 2001.

[36] X. Liu. Diffusion models for double-ended queues with reneging in heavy traffic. *Queueing Systems*, 91(1-2):49–87, 2019.

[37] Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Operations research*, 59(4):835–846, 2011.

[38] A. Mandelbaum and P. Momčilović. Personalized queues: the customer view, via a fluid model of serving least-patient first. *Queueing Systems*, 87(1):23–53, 2017.

[39] T. Masanet and P. Moyal. Perfect sampling of stochastic matching models with reneging, 2022. ArXiv preprint arXiv:2202.09341.

[40] E. Özkan. Joint pricing and matching in ride-sharing systems. *European Journal of Operational Research*, 287(3):1149–1160, 2020.

[41] E. Özkan and A. R. Ward. Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70, 2020.

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

[43] A. L. Puha and A. R. Ward. Fluid limits for multiclass many-server queues with general reneging distributions and head-of-the-line scheduling. *Mathematics of Operations Research*, 2021.

[44] A. L. Puha and R. J. Williams. Asymptotic behavior of a critical fluid model for a processor sharing queue via relative entropy. *Stochastic Systems*, 6(2):251–300, 2016.

[45] M. Remerova, J. Reed, and B. Zwart. Fluid limits for bandwidth-sharing networks with rate constraints. *Mathematics of Operations Research*, 39(3):746–774, 2014.

[46] S. A. Zenios. Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems*, 31(3-4):239–251, 1999.

[47] J. Zhang. Fluid models of many-server queues with abandonment. *Queueing Systems*, 73(2):147–193, 2013.

[48] J. Zhang, J. Dai, and B. Zwart. Law of large number limits of limited processor-sharing queues. *Mathematics of Operations Research*, 34(4):937–970, 2009.

[49] Y. Zhong, A. L. Puha, and A. R. Ward. Asymptotically optimal idling in the GI/GI/$n$+GI queue. *Operations Research Letters*, 50(3):362–369, 2022.

[50] M. Zubeldia, P. J. Jhunjhunwala, and S. T. Maguluri. Matching queues with abandonments in quantum switches: Stability and throughput analysis, 2022. ArXiv preprint arXiv:2209.12324.

[51] A. W. Zuñiga. Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions. *Stochastic Processes and their Applications*, 124(3):1436–1468, 2014.