Fluid Model of A Many-Server Queueing Network with Abandonment and Markovian Routing

WEINING KANG AND GUODONG PANG

Abstract. This paper studies a fluid model for a non-Markovian many-server queueing network with abandonment, where externally arrived and internally routed customers are served under the non-idling global First-Come-First-Serve (FCFS) discipline at each station of many parallel servers. The routing follows a Markovian mechanism. Externally arrived and internally routed customers in each queue may have different service time distributions, as well as different patience time distributions, and all these distributions may depend on the station. The fluid model dynamics is described by the fluid contents of externally arrived customers and internally routed customers in each queue (both waiting and receiving service) and a set of four measure-valued processes, tracking the amount of service time each externally arrived customer in service has received, the amount of service time each internally routed customer in service has received, the waiting times of externally arrived customers and the waiting times of internally routed customers in queue. Under mild conditions on the service and patience time distributions, we prove the existence and uniqueness of a solution to the fluid model equations. We then characterize the invariant states of this fluid model when the arrival rates are constant. We also establish the convergence of the properly scaled stochastic evolution dynamics to the fluid model.

1. Introduction

We study a fluid model for a non-Markovian many-server queueing network model with customer abandonment and Markovian routing. In the network model, there are a fixed number of service stations, each of which has either finitely or infinitely many parallel servers, a single queue and its own designated customer class. Customers enter the system at a service station, and receive service immediately if there is a free server at the station, and join the queue at the station otherwise. Our network model allows for customer internal routing, that is, upon service completion, a customer is immediately routed to one of the service stations for another service or leaves the system following a Markovian routing mechanism, independent of other customers. The service-time distributions of externally arrived and internally routed customers may be different at each station and also station dependent. Externally arrived and internally routed customers at each service station, divided as two separate classes of customers (due to possibly different service time distributions), are served in the non-idling, global First-Come-First-Serve (FCFS) discipline. Externally arrived and internally routed customers can be out of patience and leave the system (without reentry) when they are waiting in the queue before receiving service and their patience-time distributions are possibly different at each service station and also station dependent.

This model has many interesting applications in customer contact centers and patient flow analysis. In particular, empirical analysis shows that the service and patience times of customers' first visit and reentrants can be very different, see [20] and [48]. Thus, it is important to understand the system dynamics of this network model with this feature of different service-time distributions and different patience-time distributions of externally arrived and internally routed customers in each service station. However, this causes technical difficulties in the exact analysis as well as

Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250

Department of Computational Applied Mathematics and Operations Research, George R. Brown School of Engineering and Computing, Rice University, 6100 Main St., Houston, TX 77005

E-mail addresses: wkang@umbc.edu, gdpang@rice.edu.

asymptotically. In the many-server regime, even for the Markovian model with exponential service and patience times at each station, the conventional approach in [38, 33] does not work.

By adapting the measure-valued stochastic-process framework developed by Kaspi and Ramanan [18] and Kang and Ramanan [17, 16], we can obtain fluid approximations of the system dynamics in the many-server regime, where the external arrival rates (possibly time-varying) at each service station grows proportionally to the total number of servers from those service stations in the system as it increases to infinity. The focus of this paper is on the analysis of the fluid model equations, in particular, the existence and uniqueness of the solution to the fluid model, as well as the invariant state of the fluid model when the arrival rate is constant. We also establish the convergence of the properly scaled stochastic processes describing the network dynamics to the fluid model equations, by extending the arguments in [18, 17, 16].

In the fluid model (see Definition 2.1), we use a set of four measure-valued processes together with two processes counting the total number of externally arrived customers and the total number of internally routed customers at each service station to describe the system evolution dynamics. Here, two measure-valued processes keep track of the amount of service time each externally arrived customer in service has received and that of each internally routed customer in service has received, and the other two measure-valued processes keep track of the waiting times of externally arrived and internally routed customers in the queue, respectively. It is critical for our analysis to use two measure-valued processes to describe the dynamics of externally arrived and internally routed customers in each queue separately, for both the service dynamics and the waiting dynamics. This is because a single measure-valued process cannot simultaneously characterize the different service or waiting (impatient) behaviors of externally arrived and internally routed customers. The previous works [18, 17, 16] for one many-server queue with or without abandonment use only one measure-valued process for the service or waiting (impatient) behaviors. The differentiation of the service and patience-time distributions among the externally arrived and internally routed customers distinguishes our work from the existing literature, and also causes substantial challenges in the analysis.

The main results in this paper are the existence and uniqueness of a solution to the fluid model (Theorem 3.3), and the characterization of the invariant states of the fluid model (Theorem 4.4). The proof of uniqueness of a fluid model solution and characterization of the invariant states of the fluid model are much more involved than the single service station setting [17]. One complication is that all the service stations are linked together due to customers' internal routing, and the analysis of one service station inevitably involves analyzing other service stations at the same time. Moreover, the service-time and patience-time differentiation of externally arrived and internally routed customers adds more complication than the single service station setting [17], since a service station needs to serve two classes of impatient customers in a global FCFS discipline instead of just a single class of impatient customers in [17]. The two measure-valued processes describing the impatience dynamics of externally arrived and internally routed customers in a single FCFS queue at each service station are not only directly linked by the waiting-time process of the head-of-line customer, who can be either an externally arrived or internally routed customer, but also indirectly linked due to Markovian internal routing. Unlike fluid limits of single-server queues and networks in the conventional regime, where a Skorohod mapping can be identified to show the uniqueness of its solution and the convergence (due to the server idleness), the fluid limits for many-server queues and networks in the FCFS regimes do not have reflections and cannot be put in the framework of Skorohod problems. Thus, new arguments are needed in the proofs of existence and uniqueness of a solution to the fluid equations and the characterization of its invariant state to address the complications from the interconnection of service stations as well as from service-time and patience-time differentiation. A recent work in [14] studies the fluid model for multiclass many-server queues under the global FCFS discipline. Although in our model, each station appears to have two separate classes of (externally arrived and internally routed) customers in the global

FCFS discipline as in [14], due to the Markovian internal routing, the arrivals of internal customers at each service station are coming from the customer departures at each service station, which, in turn, depend on the arrivals of customers at each service station in a highly nonlinear way. So the existence and uniqueness results and techniques in [14] cannot be directly applied. We overcome this difficulty by adapting the sensitivity argument in [14] and using a localization argument. These methods may turn out to be useful for the study of existence and uniqueness of fluid models for other FCFS service networks. For completeness, we also establish the convergence of fluid scaled state descriptor for the network dynamics consisting of the two processes counting the numbers of externally arrived and internally routed customers, respectively, and the four measure-valued processes to the fluid model. We adapt the approach in [17] for the single class, single service station model to accommodate the multiclass, multiple service stations nature in our model. Our adapted argument also needs to address the complications due to the Markovian routing that does not exist in [17] (see, for example, Lemma 5.3).

Literature Review. Many-server queues with abandonment and their networks have been extensively used to model large-scale service systems, for example, customer contact centers and patient flows in hospitals; see [11], [10], [36], [12], [21], [1] and [43] and references therein. There is a vast literature on Markovian many-server queueing (network) models with abandonment. We refer the readers to the above cited papers for a complete review of them. Empirical study of call centers and patient flows have shown that customers' service and patience times are usually non-exponential; see, e.g., [4], [37], [1] and [43]. Thus, it is significant that stochastic models for these systems capture the realistic feature of non-exponential service and patience time distributions. There has been substantial development in the recent years on non-Markovian many-server queueing (network) models. Here we review those mostly relevant to our work. (i) For non-Markovian many-server queues, we refer to [46, 18, 16, 17, 49, 23, 24, 25, 26, 15, 45, 30, 44, 13, 35, 29, 2] for fluid models using measure-valued and two-parameter processes tracking elapsed or residual times and [19, 41, 5, 6, 34, 8] for approximations in the Halfin-Whitt regime. See also the recent surveys in [45, 7, 22]. (ii) There is very limited research on non-Markovian many-server queuing network models with abandonment. Atar et al. [3] generalize the measure-valued process approach in [18, 16] to study a multiclass non-Markovian many-server model with abandonment, in which customers are served according to a non-preemptive priority policy. Liu and Whitt [27, 28] study a fluid network model for a non-Markovian open queueing network of many-server queues, where all model elements are time-varying. They generalize their algorithm in [23] to this time-varying fluid network model. Their model is closest to ours, but they do not consider the differentiation of service and patience times of external and internal customers. This model difference causes nontrivial analytical difficulties. For example, in their model, the invariant state is characterized via that of the $G/GI/s+GI$ model [46] and the aggregate arrival rates (satisfying a fixed point equation); however, that approach does not work for our model with differentiated service and patience time distributions for externally arrived and internally routed customers. We also refer to [42, 31, 40, 32] on studies of dynamic scheduling and routing in multiclass or multi-pool many-server queues.

1.1. **Organization of the Paper.** The notation used in this paper is the same as in [18, 17, 16], so we give a brief description of notation in §1.2. We describe the model primitives in §2.1 and present the fluid model in §2.2. We prove the existence and uniqueness of a solution to the fluid equations in §3. The characterization of its invariant state and its proof are given in §4. In §5, we establish the convergence of the properly scaled process to the fluid model equations.

1.2. **Notation.** The following notation will be used throughout the paper. $\mathbb{N}$, $\mathbb{R}$ ($\mathbb{R}_+$), $\mathbb{Z}$ ($\mathbb{Z}_+$) represent the set of strictly positive integers, real numbers (non-negative), integers (non-negative), respectively. For $a, b \in \mathbb{R}$, $a \vee b = max\{a, b\}$, $a \wedge b = \min\{a, b\}$ and $a^+ = a \vee 0$. For a set B, $\mathbb{1}_B$ denotes the indicator function of the set B. For a square matrix P, P' denotes its transpose, P^{-1} denotes its inverse if P is invertible, and P^n denotes its nth power. For any metric space E, $\mathcal{C}_b(E)$

and $\mathcal{C}_c(E)$ are, respectively, the space of bounded, continuous functions and the space of continuous real-valued functions with compact support defined on E, while $\mathcal{C}^1(E)$ is the space of real-valued, once continuously differentiable functions on E, and $\mathcal{C}^1_c(E)$ is the subspace of functions in $\mathcal{C}^1(E)$ that have compact support. The subspace of functions in $\mathcal{C}^1(E)$ that, together with their first derivatives, are bounded, will be denoted by $\mathcal{C}^1_b(E)$. For $\varphi : E \to \mathbb{R}$, let $\|\varphi\|_\infty \doteq \sup_{x\in E} |\varphi(x)|$ and supp(φ) be the support of φ. For $H \leq \infty$, let $\mathcal{L}^1[0, H)$ and $\mathcal{L}^1_{loc}[0, H)$, respectively, represent the spaces of integrable and locally integrable functions on $[0, H)$, where a locally integrable function f on $[0, H)$ is a measurable function on $[0, H)$ that satisfies $\int_{[0,a]} |f(x)|dx < \infty$ for all $a < H$. The constant functions $f \equiv 1$ and $f \equiv 0$ will be represented by the symbols $\mathbf{1}$ and $\mathbf{0}$, respectively. For any càdlàg function $f : [0, \infty) \to \mathbb{R}$, $\|f\|_T \doteq \sup_{s\in[0,T]} |f(s)|$ for every $T < \infty$. Given a non-decreasing, right continuous function f having left limits on $[0, \infty)$, f^{-1} denotes the left continuous inverse function of f: $f^{-1}(y) = \inf\{x \geq 0 : f(x) \geq y\}$ with the convention that infimum over an empty set is ∞. For each differentiable function f defined on $\mathbb{R}$, f' denotes the first derivative of f. For each function $f(t, x)$ defined on $\mathbb{R} \times \mathbb{R}^n$, f_t denotes the partial derivative of f with respect to t and f_x denotes the partial derivative of f with respect to $x \in \mathbb{R}$.

The space of Radon measures on a Polish space E, endowed with the Borel σ-algebra, is denoted by $\mathcal{M}(E)$, while $\mathcal{M}_+(E)$ and $\mathcal{M}_F(E)$ are, respectively, the subspaces of non-negative, finite non-negative measures in $\mathcal{M}(E)$. $\mathcal{M}(E)$ ($\mathcal{M}_+(E)$) and $\mathcal{M}_F(E)$, endowed with the vague and weak topologies [39, 47], respectively, are Polish spaces. The symbol δ_x denotes the measure with unit mass at the point x. We will also use $\mathbf{0}$ to denote the identically zero Radon measure on E. When E is an interval, say $[0, H)$, for notational conciseness, we will often write $\mathcal{M}[0, H)$ instead of $\mathcal{M}([0, H))$. We say a measure μ is continuous at $x \in [0, H)$ if and only if $\mu(\{x\}) = 0$ and μ is continuous on $[0, H)$ if μ is continuous at each $x \in [0, H)$. When $E = [0, H)$ and $E = [0, H) \times \mathbb{R}_+$, for some $H \in (0, \infty]$. we will usually use f to denote generic functions on $[0, H)$ and φ to denote generic functions on $[0, H)\times\mathbb{R}_+$. For any Borel measurable function $f : [0, H) \to \mathbb{R}$ that is integrable with respect to $\xi \in \mathcal{M}[0, H)$, we often use the short-hand notation $\langle f, \xi\rangle \doteq \int_{[0,H)} f(x)\,\xi(dx)$. For each measure μ on $[0, \infty)$, let $F^\mu(x) \doteq \mu[0, x]$ for each $x \in [0, \infty)$. Let $\mathcal{I}_{\mathbb{R}_+}[0, \infty)$ be the subset of non-decreasing functions $f \in \mathcal{D}_{\mathbb{R}_+}[0, \infty)$ with $f(0) = 0$.

2. The Fluid Model

2.1. **The Network Model and Primitive Data.** Consider a system with K ($1 \leq K < \infty$) service stations and each service station has its own designated customer class. Thus, there are K customer classes for the entire system. Let $\mathcal{K} \doteq \{1, \cdots, K\}$. For each $k \in \mathcal{K}$, customers of class k are served by the kth service station with $N_k \in [1, \infty]$ identical servers, in which arriving customers are served in a non-idling, FCFS manner, that is, a newly arriving customer immediately enters service if there are any idle servers or, if all servers are busy, then the customer joins the end of the queue, and the customer at the head of the queue (if one is present) enters service as soon as a server becomes free. Note that we allow the possibility that a service station may have infinitely many identical servers. Let N be a positive integer. We assume that, for each $k \in \mathcal{K}$, $N_k = \lfloor s_k N \rfloor$, where s_k is a fixed constant in $(0, \infty]$ independent of N. Note that for each $k \in \mathcal{K}$, $N_k = \infty$ if $s_k = \infty$ and $N_k/N \to s_k$ as $N \to \infty$. (Since we focus on the fluid model in this paper, the scaling parameter N will not appear in the fluid model described in §2.2.)

External Arrivals. We assume that there exists a K-dimensional cumulative external arrival process $E^{(N)}$ such that for each $k \in \mathcal{K}$ and $t > 0$, $E^{(N)}_k(t)$ represents the total number of customers of class k that have arrived into the system from outside during the time interval $(0, t]$. We assume that $E^{(N)}_k$ is a non-decreasing, pure jump process with $E^{(N)}_k(0) = 0$ and a.s., for each $t \in [0, \infty)$, $E^{(N)}_k(t) < \infty$ and $E^{(N)}_k(t) - E^{(N)}_k(t-) \in \{0, 1\}$. Also, for each $k \in \mathcal{K}$, let $\mathcal{E}^{(N)}_k$ be an a.s. $\mathbb{Z}_+$-valued random variable that represents the number of customers of class k that have entered

the system by time zero. The set of random variables $\{\mathcal{E}_k^{(N)}, k \in \mathcal{K}\}$ is only used for bookkeeping purposes to keep track of the indices of customers. We assume that there exist deterministic non-decreasing functions $\bar{E}_k$ in $\mathcal{D}_{\mathbb{R}_+}[0,\infty)$ with $\overline{E}_k(0) = 0$ such that $N^{-1}E_k^{(N)} \Rightarrow \overline{E}_k$ as $N \to \infty$.

Service Times. For each $k \in \mathcal{K}$, the customers of class k are either coming externally from the outside of the system, or coming internally from one of the service stations upon service completion due to internal routing. We shall call the first type of customers as *external customers* and the second type of customers as *internal customers*. We assume that for each $k \in \mathcal{K}$, there exists two sequences of i.i.d. random variables $\{v_i^{1,k},\ i \in \mathbb{Z}\}$ with the cumulative distribution function (c.d.f.) $G_{1,k}^s$ on $[0,\infty)$ and $\{v_i^{2,k},\ i \in \mathbb{Z}\}$, with the c.d.f. $G_{2,k}^s$ on $[0,\infty)$. For each $k \in \mathcal{K}$ and $i \in \mathbb{N}$, $v_i^{1,k}$ (resp. $v_i^{2,k}$) represents the service requirement of the ith external (resp. internal) customer of class k to enter the system after time zero, while $\{v_i^{1,k},\ i \in -\mathbb{N} \cup \{0\}\}$ (resp. $\{v_i^{2,k},\ i \in -\mathbb{N} \cup \{0\}\}$) represents the service requirements of external (resp. internal) customers of class k that arrived by time zero (if such customers exist), ordered according to their arrival times (by time zero). We assume that $G_{1,k}^s$ and $G_{2,k}^s$ have densities $g_{1,k}^s$ and $g_{2,k}^s$, respectively. For each $j = 1, 2$, let

$$H_{j,k}^s \doteq \sup\{x \in [0,\infty) : G_{j,k}^s(x) < 1\}.$$

Then $H_{1,k}^s$ and $H_{2,k}^s$ denote, respectively, the right end of the support of $g_{1,k}^s$ and $g_{2,k}^s$. We assume that the two service time distributions $G_{1,k}^s$ and $G_{2,k}^s$ have positive finite means, that is,

$$m_{1,k}^s \doteq \int_{[0,H_{1,k}^s)} (1 - G_{1,k}^s(x))dx \in (0,\infty) \text{ and } m_{2,k}^s \doteq \int_{[0,H_{2,k}^s)} (1 - G_{2,k}^s(x))dx \in (0,\infty). \tag{2.1}$$

Markovian Routing. We assume Markovian routing, described as follows. Let $e_1, \cdots, e_K$ be the K unit coordinate vectors in $\mathbb{R}^K$ and e_0 be the K-dimensional vector of all zeros. For each $k \in \mathcal{K}$, $\{\phi^{1,k}(i), i \in \mathbb{Z}\}$ and $\{\phi^{2,k}(i), i \in \mathbb{Z}\}$ are two sequences of i.i.d. routing vectors where for $j = 1, 2$ and $i \in \mathbb{Z}$, $\phi^{j,k}(i)$ takes values in the set $\{e_0, e_1, \cdots, e_K\}$. For each $k \in \mathcal{K}$ and $i \in \mathbb{Z}$, the ith external (resp. internal) customer of class k to depart from the kth service station is next routed to class l if $\phi^{1,k}(i) = e_l$ (resp. $\phi^{2,k}(i) = e_l$) for some $l \in \mathcal{K}$, or it leaves the system if $\phi^{1,k}(i) = e_0$ (resp. $\phi^{2,k}(i) = e_0$). Let P be a $K \times K$ matrix such that for each $k, l \in \mathcal{K}$ and $i \in \mathbb{Z}$,

$$P_{kl} = \mathbb{P}(\phi^{1,k}(i) = e_l) = \mathbb{P}(\phi^{2,k}(i) = e_l).$$

The matrix P is called the routing matrix. Let $P_{k0} \doteq 1 - \sum_{l \in \mathcal{K}} P_{kl}$ for each $k \in \mathcal{K}$. We assume that P satisfies the conditions that $I - P'$ is invertible and

$$H \doteq (I - P')^{-1} = I + P' + (P')^2 + (P')^3 + \cdots.$$

Note that the matrix H has non-negative entries.

Reneging. It is assumed that customers are impatient, and that a customer reneges from the queue as soon as the amount of time it has spent in queue reaches its patience limit. Customers do not renege once they have entered service. We assume that external customers and internal customers in queue may have different patience time distributions. For each $k \in \mathcal{K}$, the patience times of external customers of class k are given by an i.i.d. sequence, $\{r_i^{1,k},\ i \in \mathbb{Z}\}$, with a common c.d.f. $G_{1,k}^r$ on $[0,\infty]$, where for each $i \in \mathbb{N}$, $r_i^{1,k}$ represents the patience time of the ith external customer of class k to enter the system after time zero, while $\{r_i^{1,k},\ i \in -\mathbb{N} \cup \{0\}\}$ represents the patience times of external customers of class k that entered the system by time zero, ordered according to their arrival times (by time zero). For each $k \in \mathcal{K}$, the patience times of internal customers of class k are given by another i.i.d. sequence, $\{r_i^{2,k},\ i \in \mathbb{Z}\}$, with a common c.d.f. $G_{2,k}^r$ on $[0,\infty]$, where for each $i \in \mathbb{N}$, $r_i^{2,k}$ represents the patience time of the ith internal customer of class k to reenter the system after time zero, while $\{r_i^{2,k},\ i \in -\mathbb{N} \cup \{0\}\}$ represents the patience times of internal customers of class k arrived by time zero, ordered according to their reentering

times (by time zero). We assume that for $j = 1, 2$, $G^r_{j,k}$, restricted on $[0, \infty)$, has density $g^r_{j,k}$. For $j = 1, 2$, let $H^r_{j,k} \doteq \sup\{x \in [0, \infty) : G^r_{j,k}(x) < 1\}$ denote, respectively, the right end of the support of $g^r_{j,k}$. For each $j = 1, 2$, the means of the two patience time distributions $G^r_{1,k}$ and $G^r_{2,k}$ are denoted by

$$m^r_{1,k} \doteq \int_{[0,H^r_{1,k})} (1 - G^r_{1,k}(x))dx \in [0, \infty] \text{ and } m^r_{2,k} \doteq \int_{[0,H^r_{2,k})} (1 - G^r_{2,k}(x))dx \in [0, \infty]. \tag{2.2}$$

Independence We assume that the cumulative external arrival processes $E^{(N)}_k$, $k \in \mathcal{K}$, the sequences of service requirements $\{v^{j,k}_i, i \in \mathbb{Z}\}$, $j = 1, 2$, $k \in \mathcal{K}$, the sequences of patience times $\{r^{j,k}_i, i \in \mathbb{Z}\}$, $j = 1, 2$, $k \in \mathcal{K}$, and the sequences of feedback vectors $\{\phi^{j,k}(i), i \in \mathbb{N}\}$, $j = 1, 2$, $k \in \mathcal{K}$ are mutually independent.

2.2. **Fluid Model Equations.** Define

$$\mathcal{S}_0 \doteq \left\{ \begin{array}{c} (e, x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2) \in \mathcal{I}_{\mathbb{R}_+}[0, \infty)^K \times \mathbb{R}^K_+ \times \mathbb{R}^K_+ \times \Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H^s_{1,k}) \\ \times \Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H^s_{2,k}) \times \Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H^r_{1,k}) \times \Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H^r_{2,k}) : \\ s_k - \big(\langle \mathbf{1}, \nu^{1,k}\rangle + \langle \mathbf{1}, \nu^{2,k}\rangle\big) = [s_k - x^1_k - x^2_k]^+, \\ [x^1_k + x^2_k - s_k]^+ \le \langle \mathbf{1}, \eta^{1,k}_0\rangle + \langle \mathbf{1}, \eta^{2,k}_0\rangle \end{array} \right\}, \tag{2.3}$$

Recall that $N^k/N \to s_k$ as $N \to \infty$. The set $\mathcal{S}_0$ serves as the space of possible input data for the fluid model equations. In order to state the definition of fluid model equations, for each $j = 1, 2$ and $k \in \mathcal{K}$, define the hazard rate functions of $G^r_{j,k}$ and $G^s_{j,k}$ in the usual manner:

$$h^r_{j,k}(x) \doteq \frac{g^r_{j,k}(x)}{1 - G^r_{j,k}(x)},\ x \in [0, H^r_{j,k}), \text{ and } h^s_{j,k}(x) \doteq \frac{g^s_{j,k}(x)}{1 - G^s_{j,k}(x)},\ x \in [0, H^s_{j,k}). \tag{2.4}$$

It is easy to verify that $h^r_{j,k} \in \mathcal{L}^1_{loc}[0, H^r_{j,k})$ and $h^s_{j,k} \in \mathcal{L}^1_{loc}[0, H^s_{j,k})$ for $j = 1, 2$.

Definition 2.1. *The càdlàg function $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ defined on $\mathbb{R}_+$ such that $\overline{X}^1 = (\overline{X}^1_k, k \in \mathcal{K}) \in \mathbb{R}^K_+$, $\overline{X}^2 = (\overline{X}^2_k, k \in \mathcal{K}) \in \mathbb{R}^K_+$, $\overline{\nu}^1 = (\overline{\nu}^{1,k}, k \in \mathcal{K}) \in \Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H^s_{1,k})$, $\overline{\nu}^2 = (\overline{\nu}^{2,k}, k \in \mathcal{K}) \in \Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H^s_{2,k})$, $\overline{\eta}^1 = (\overline{\eta}^{1,k}, k \in \mathcal{K}) \in \Pi_{k\in\mathcal{K}}\mathcal{M}_+[0, H^r_{1,k})$, and $\overline{\eta}^2 = (\overline{\eta}^{2,k}, k \in \mathcal{K}) \in \Pi_{k\in\mathcal{K}}\mathcal{M}_+[0, H^r_{2,k})$ is said to solve the fluid model equations associated with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$ and the hazard rate functions $h^r_{j,k}$ and $h^s_{j,k}$, $j = 1, 2$ and $k \in \mathcal{K}$, if and only if for every $t \in [0, \infty)$, $j = 1, 2$ and $k \in \mathcal{K}$,*

$$\int_0^t \langle h^r_{j,k}, \overline{\eta}^{j,k}_u\rangle\, du < \infty, \qquad \int_0^t \langle h^s_{j,k}, \overline{\nu}^{j,k}_u\rangle\, du < \infty, \tag{2.5}$$

and the following relations are satisfied: for every $f \in \mathcal{C}_b(\mathbb{R}_+)$,

$$\begin{aligned} \int_{[0,H^s_{1,k})} f(x)\, \overline{\nu}^{1,k}_t(dx) &= \int_{[0,H^s_{1,k})} f(x+t) \frac{1 - G^s_{1,k}(x+t)}{1 - G^s_{1,k}(x)}\, \overline{\nu}^{1,k}_0(dx) \\ &\quad + \int_{[0,t]} f(t-s)(1 - G^s_{1,k}(t-s))\, d\overline{L}^1_k(s), \end{aligned} \tag{2.6}$$

where

$$\overline{L}^1_k(t) = \langle \mathbf{1}, \overline{\nu}^{1,k}_t\rangle - \langle \mathbf{1}, \overline{\nu}^{1,k}_0\rangle + \int_0^t \langle h^s_{1,k}, \overline{\nu}^{1,k}_u\rangle\, du, \tag{2.7}$$

$$\int_{[0,H^s_{2,k})} f(x)\, \overline{\nu}^{2,k}_t(dx) = \int_{[0,H^s_{2,k})} f(x+t) \frac{1 - G^s_{2,k}(x+t)}{1 - G^s_{2,k}(x)}\, \overline{\nu}^{2,k}_0(dx)$$

$$+\int_{[0,t]} f(t-s)(1-G^s_{2,k}(t-s))\,d\overline{L}^2_k(s), \tag{2.8}$$

where

$$\overline{L}^2_k(t) = \langle \mathbf{1}, \overline{\nu}^{2,k}_t\rangle - \langle \mathbf{1}, \overline{\nu}^{2,k}_0\rangle + \int_0^t \langle h^s_{2,k}, \overline{\nu}^{2,k}_u\rangle\,du, \tag{2.9}$$

$$\begin{aligned}\int_{[0,H^r_{1,k})} f(x)\,\overline{\eta}^{1,k}_t(dx) &= \int_{[0,H^r_{1,k})} f(x+t)\frac{1-G^r_{1,k}(x+t)}{1-G^r_{1,k}(x)}\,\overline{\eta}^{1,k}_0(dx)\\ &\quad+\int_{[0,t]} f(t-s)(1-G^r_{1,k}(t-s))\,d\overline{E}_k(s),\end{aligned} \tag{2.10}$$

$$\begin{aligned}\int_{[0,H^r_{2,k})} f(x)\,\overline{\eta}^{2,k}_t(dx) &= \int_{[0,H^r_{2,k})} f(x+t)\frac{1-G^r_{2,k}(x+t)}{1-G^r_{2,k}(x)}\,\overline{\eta}^{2,k}_0(dx)\\ &\quad+\int_0^t f(t-s)(1-G^r_{2,k}(t-s))\,d\overline{I}_k(s),\end{aligned} \tag{2.11}$$

where

$$\overline{I}_k(t) = \sum_{l\in\mathcal{K}} P_{lk}\int_0^t \left(\langle h^s_{1,l}, \overline{\nu}^{1,l}_u\rangle + \langle h^s_{2,l}, \overline{\nu}^{2,l}_u\rangle\right)\,du, \tag{2.12}$$

$$\overline{Q}^1_k(t) = \overline{X}^1_k(t) - \langle \mathbf{1}, \overline{\nu}^{1,k}_t\rangle = \langle \mathbf{1}_{[0,\overline{\chi}_k(t)]}, \overline{\eta}^{1,k}_t\rangle, \tag{2.13}$$

$$\overline{Q}^2_k(t) = \overline{X}^2_k(t) - \langle \mathbf{1}, \overline{\nu}^{2,k}_t\rangle = \langle \mathbf{1}_{[0,\overline{\chi}_k(t)]}, \overline{\eta}^{2,k}_t\rangle, \tag{2.14}$$

$$\overline{R}^1_k(t) = \int_0^t \left(\int_{[0,H^r_{1,k})} \mathbf{1}_{[0,\overline{\chi}_k(s)]}(u) h^r_{1,k}(u)\overline{\eta}^{1,k}_s(du)\right) ds, \tag{2.15}$$

$$\overline{R}^2_k(t) = \int_0^t \left(\int_{[0,H^r_{2,k})} \mathbf{1}_{[0,\overline{\chi}_k(s)]}(u) h^r_{2,k}(u)\overline{\eta}^{2,k}_s(du)\right) ds, \tag{2.16}$$

where $\overline{\chi}_k(s) = (F^{\overline{\eta}^k_s})^{-1}(\overline{Q}^1_k(s) + \overline{Q}^2_k(s))$ *and* $\overline{\eta}^k_s \doteq \overline{\eta}^{1,k}_s + \overline{\eta}^{2,k}_s$,

$$\overline{X}^1_k(t) = \overline{X}^1_k(0) + \overline{E}_k(t) - \int_0^t \langle h^s_{1,k}, \overline{\nu}^{1,k}_u\rangle\,du - \overline{R}^1_k(t), \tag{2.17}$$

$$\overline{X}^2_k(t) = \overline{X}^2_k(0) + \overline{I}_k(t) - \int_0^t \langle h^s_{2,k}, \overline{\nu}^{2,k}_u\rangle\,du - \overline{R}^2_k(t), \tag{2.18}$$

and

$$s_k - \langle \mathbf{1}, \overline{\nu}^{1,k}_t\rangle - \langle \mathbf{1}, \overline{\nu}^{2,k}_t\rangle = [s_k - \overline{X}^1_k(t) - \overline{X}^2_k(t)]^+. \tag{2.19}$$

It immediately follows from (2.13), (2.14) and (2.19) that for each $t \in [0,\infty)$,

$$\overline{Q}^1_k(t) + \overline{Q}^2_k(t) = [\overline{X}^1_k(t) + \overline{X}^2_k(t) - s_k]^+, \tag{2.20}$$

and from (2.7),(2.9), (2.13), (2.14), (2.17) and (2.17), for every $t \in [0,\infty)$ and $k \in \mathcal{K}$,

$$\overline{Q}^1_k(0) + \overline{E}_k(t) = \overline{Q}^1_k(t) + \overline{L}^1_k(t) + \overline{R}^2_k(t). \tag{2.21}$$

$$\overline{Q}^2_k(0) + \overline{I}_k(t) = \overline{Q}^2_k(t) + \overline{L}^2_k(t) + \overline{R}^2_k(t). \tag{2.22}$$

Remark 2.2. *From the definition of the fluid model equations in Definition 2.1, it is ready to verify that if* $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ *is a solution to the fluid model equations associated with the input data* $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0)$*, then for each* $k \in \mathcal{K}$, $(\overline{X}^1_k, \overline{X}^2_k, \overline{\nu}^{1,k}, \overline{\nu}^{2,k}, \overline{\eta}^{1,k}, \overline{\eta}^{2,k})$ *is a solution to the fluid model equations associated with the input data* $(\overline{E}_k, \overline{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$

as in Definition 2.1 of [14]. *Intuitively, at service station $k \in \mathcal{K}$, there are two classes of customers: externally arrived customers with the arrival process $\overline{E}_k$ and internally routed customers with the arrival process $\overline{I}_k$, and the two classes of customers have differentiated service time distributions $G^s_{1,k}$ and $G^s_{2,k}$ and differentiated patience time distributions $G^r_{1,k}$ and $G^r_{2,k}$, respectively. The service discipline at each service station k is the global first-come-first-serve discipline (global FCFS), that is, a server will serve the oldest customer waiting in queue irrespective of its customer class at the moment when it becomes available, and non-idling, that is, no server will idle whenever there is a customer of any class in queue.*

Remark 2.3. *It follows from (2.12), (2.15) and (2.16) that for each $k \in \mathcal{K}$, processes $\overline{I}_k$, $\overline{R}^1_k$ and $\overline{R}^2_k$ are continuous. If the process $\overline{E}_k$ is also continuous for each $k \in \mathcal{K}$, the relations in (2.17) and (2.18) implies that the processes $\overline{X}^1_k$ and $\overline{X}^2_k$ and then $\overline{Q}^1_k$ and $\overline{Q}^2_k$ by (2.13) and (2.14) are continuous. Moreover, for each $k \in \mathcal{K}$, the continuity of $\overline{E}_k$ also implies that $\overline{L}^1_k$ and $\overline{L}^2_k$ are continuous by (2.21) and (2.22).*

Remark 2.4. *For each $k \in \mathcal{K}$, if $s_k = \infty$, the non-idling condition (2.19) holds automatically and in this case, by (2.20), $\overline{Q}^1_k(t) + \overline{Q}^2_k(t) = 0$ for all $t \geq 0$, and then $\overline{\chi}_k(t) = \overline{R}^1_k(t) = \overline{R}^2_k(t) = 0$ for all $t \geq 0$ by (2.15) and (2.16).*

We close this section with a simple result on the action of time-shifts on solutions to the fluid model equations. For this, we need the following notation: for any $t \in \mathbb{R}_+$, $k \in \mathcal{K}$ and $j = 1, 2$,

$$\overline{E}^{[t]} \doteq \overline{E}(t+\cdot) - \overline{E}(t), \qquad \overline{L}^{[t],j} \doteq \overline{L}^j(t+\cdot) - \overline{L}^j(t), \qquad \overline{X}^{[t],j} \doteq \overline{X}^j(t+\cdot), \qquad \overline{\nu}^{[t],j} \doteq \overline{\nu}^j_{t+\cdot},$$

$$\overline{R}^{[t],j} \doteq \overline{R}^j(t+\cdot) - \overline{R}^j(t), \qquad \overline{\eta}^{[t],j} \doteq \overline{\eta}^j_{t+\cdot}, \qquad \overline{Q}^{[t],j} \doteq \overline{Q}^j(t+\cdot).$$

Lemma 2.5. *Suppose the càdlàg function $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ defined on $\mathbb{R}_+$ solves the fluid model equations associated with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$. Then the time-shifted function $(\overline{X}^{[t],1}, \overline{X}^{[t],2}, \overline{\nu}^{[t],1}, \overline{\nu}^{[t],2}, \overline{\eta}^{[t],1}, \overline{\eta}^{[t],2})$ solves the fluid model equations associated with the input data $(\overline{E}^{[t]}, \overline{X}^1(t), \overline{X}^2(t), \overline{\nu}^1_t, \overline{\nu}^2_t, \overline{\eta}^1_t, \overline{\eta}^2_t) \in \mathcal{S}_0$, where $\overline{L}^{[t],1}, \overline{L}^{[t],2}, \overline{R}^{[t],1}, \overline{R}^{[t],2}, \overline{Q}^{[t],1}, \overline{Q}^{[t],2}$ are the corresponding processes that satisfy (2.7),(2.9), (2.15), (2.16),(2.13),(2.14), with $\overline{X}^{[t],1}$, $\overline{X}^{[t],2}$, $\overline{\nu}^{[t],1}$, $\overline{\nu}^{[t],2}$, $\overline{\eta}^{[t],1}$, $\overline{\eta}^{[t],2}$ in place of $\overline{X}^1$, $\overline{X}^2$, $\overline{\nu}^1$, $\overline{\nu}^2$, $\overline{\eta}^1$, $\overline{\eta}^2$.*

The proof of the lemma just involves a rewriting of the fluid equations, and is thus omitted.

3. Well-posedness of Solutions to the Fluid Model Equations

In this section, we establish the existence and uniqueness of solutions to the fluid model equations with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$ that satisfies the following three assumptions.

Assumption 3.1. *The arrival process $\overline{E} = (\overline{E}_k, k \in \mathcal{K})$ is absolutely continuous with a.e. derivative $\overline{\lambda}(\cdot) = (\overline{\lambda}_k(\cdot), k \in \mathcal{K})$, for each $k \in \mathcal{K}$, $\overline{\eta}^{1,k}_0(\{x\}) = 0$ and $\overline{\eta}^{2,k}_0(\{x\}) = 0$ for all $x \in \mathbb{R}_+$, the hazard rate functions $\{h^r_{1,k}, h^r_{2,k},\ k \in \mathcal{K}\}$ of the patience time distributions $\{G^r_{1,k}, G^r_{2,k},\ k \in \mathcal{K}\}$ are a.e. locally bounded and the densities $\{g^s_{1,k}, g^s_{2,k},\ k \in \mathcal{K}\}$ of the service time distributions $\{G^s_{1,k}, G^s_{2,k},\ k \in \mathcal{K}\}$ satisfy that for each $k \in \mathcal{K}$ and $j = 1, 2$, there is an integer $q^j_k \geq 1$ such that for each $S > 0$,*

$$\int_0^S |g^s_{j,k}(s+h) - g^s_{j,k}(s)|^{q^j_k} ds \to 0 \quad \text{as} \quad h \downarrow 0. \tag{3.1}$$

Moreover, if $h^r_{j,k}$ is unbounded on $[0, H^r_{j,k})$ for some $k \in \mathcal{K}$ and some $j = 1, 2$, it is assumed that

$$\overline{\chi}_k(0) = (F^{\overline{\eta}^k_0})^{-1}\left(\left[\overline{X}^1_k(0) + \overline{X}^2_k(0) - s_k\right]^+\right) < \infty. \tag{3.2}$$

Remark 3.1. *Since the hazard rate function of any distribution is only locally integrable and never integrable over its support, then when we assume that the hazard rate functions $\{h^r_{1,k}, h^r_{2,k},\ k \in \mathcal{K}\}$ are a.e. locally bounded in Assumption 3.1, we implicitly assume that $H^r_{j,k} = \infty$ for all $k \in \mathcal{K}$ and $j = 1, 2$. The condition (3.1) on the service time densities $\{g^s_{1,k}, g^s_{2,k},\ k \in \mathcal{K}\}$ is not too restrictive. For example, if $\{g^s_{1,k}, g^s_{2,k},\ k \in \mathcal{K}\}$ are right continuous, then they satisfy (3.1) with $q^1_k = q^2_k = 1$, by a simple application of dominated convergence theorem.*

Assumption 3.2. *The service time densities $\{g^s_{1,k}, g^s_{2,k},\ k \in \mathcal{K}\}$ are right continuous on their supports and are absolutely continuous on $[0, \delta]$ for some $\delta > 0$, and for each $k \in \mathcal{K}$, one of the following two conditions holds:*

(A) *There exists some $j = 1, 2$ such that $\int_{[0,H^s_{j,k})} \frac{g^s_{j,k}(x+t)}{1-G^s_{j,k}(x)}\, \overline{\nu}^{j,k}_0(dx) > 0$ for all $t \in \mathbb{R}_+$.*

(B) *For each $j = 1, 2$, $h^s_{j,k}(x) > 0$ for each $x \in [0, H^s_{j,k})$.*

Assumption 3.3. *The densities $\{g^r_{1,k},\ g^r_{2,k},\ k \in \mathcal{K}\}$ of the patience time distributions $\{G^r_{1,k},\ G^r_{2,k},\ k \in \mathcal{K}\}$ are absolutely continuous on $[0, \delta)$ for some $\delta > 0$.*

Remark 3.2. *As discussed in Remark 2.2, if $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ is a solution to the fluid model equations associated with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0)$, then for each $k \in \mathcal{K}$, $(\overline{X}^1_k, \overline{X}^2_k, \overline{\nu}^{1,k}, \overline{\nu}^{2,k}, \overline{\eta}^{1,k}, \overline{\eta}^{2,k})$ is a solution to the fluid model equations associated with the input data $(\overline{E}_k, \overline{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$. Fix $k \in \mathcal{K}$. Since $\overline{I}_k$ is absolutely continuous with derivative $\sum_{l\in\mathcal{K}} P_{lk}\left(\langle h^s_{1,l}, \overline{\nu}^{1,l}_\cdot\rangle + \langle h^s_{2,l}, \overline{\nu}^{2,l}_\cdot\rangle\right)$ by (2.12), this and Assumption 3.1 and Assumption 3.2 restricted to $k \in \mathcal{K}$ together imply that Assumptions 3.1 and 3.2 of [14] hold for the input data $(\overline{E}_k, \overline{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$. Then by Theorems 3.6 and 3.7 of [14], it follows that there exists a unique solution to the fluid model equations in Definition 2.1 of [14] associated with the input data $(\overline{E}_k, \overline{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$, where*

$$\overline{I}_k(t) = \sum_{l\in\mathcal{K}} P_{lk} \int_0^t \left(\langle h^s_{1,l}, \overline{\nu}^{1,l}_u\rangle + \langle h^s_{2,l}, \overline{\nu}^{2,l}_u\rangle\right) du,\ t \in [0, \infty),$$

Theorem 3.3. *Under Assumptions 3.1–3.3, there exists a unique solution $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ to the fluid model equations associated with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$ in Definition 2.1.*

Proof. We first establish the uniqueness of solutions to the fluid model equations associated with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$. Suppose that there are two such solutions $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ and $(\hat{X}^1, \hat{X}^2, \hat{\nu}^1, \hat{\nu}^2, \hat{\eta}^1, \hat{\eta}^2)$ to the fluid model equations associated with the same input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$. From the discussion in Remark 3.2, for each $k \in \mathcal{K}$, $(\overline{X}^1_k, \overline{X}^2_k, \overline{\nu}^{1,k}, \overline{\nu}^{2,k}, \overline{\eta}^{1,k}, \overline{\eta}^{2,k})$ is a solution to the fluid model equations associated with the input data $(\overline{E}_k, \overline{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$ in Definition 2.1 of [14], where

$$\overline{I}_k(t) = \sum_{l\in\mathcal{K}} P_{lk} \int_0^t \left(\langle h^s_{1,l}, \overline{\nu}^{1,l}_u\rangle + \langle h^s_{2,l}, \overline{\nu}^{2,l}_u\rangle\right) du,\ t \in [0, \infty),$$

and $(\hat{X}^1_k, \hat{X}^2_k, \hat{\nu}^{1,k}, \hat{\nu}^{2,k}, \hat{\eta}^{1,k}, \hat{\eta}^{2,k})$ is a solution to the fluid model equations associated with the input data $(\overline{E}_k, \hat{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$ in Definition 2.1 of [14], where

$$\hat{I}_k(t) = \sum_{l\in\mathcal{K}} P_{lk} \int_0^t \left(\langle h^s_{1,l}, \hat{\nu}^{1,l}_u \rangle + \langle h^s_{2,l}, \hat{\nu}^{2,l}_u \rangle \right) du, \ t \in [0,\infty).$$

Since $(\overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$ is the initial state for the solution $(\hat{X}^1_k, \hat{X}^2_k, \hat{\nu}^{1,k}, \hat{\nu}^{2,k}, \hat{\eta}^{1,k}, \hat{\eta}^{2,k})$, then it follows that $\hat{\eta}^{1,k}_0 = \overline{\eta}^{1,k}_0$, $\hat{\eta}^{2,k}_0 = \overline{\eta}^{2,k}_0$, $\hat{X}^1_k(0) = \overline{X}^1_k(0)$ and $\hat{X}^2_k(0) = \overline{X}^2_k(0)$. Recall in Definition 2.1 that

$$\overline{\chi}_k(0) = (F^{\overline{\eta}^k_0})^{-1} \left(\left[\overline{X}^1_k(0) + \overline{X}^2_k(0) - s_k \right]^+ \right), \text{ where } \overline{\eta}^k_0 = \overline{\eta}^{1,k}_0 + \overline{\eta}^{2,k}_0,$$

and

$$\hat{\chi}_k(0) = (F^{\hat{\eta}^k_0})^{-1} \left(\left[\hat{X}^1_k(0) + \hat{X}^2_k(0) - s_k \right]^+ \right), \text{ where } \hat{\eta}^k_0 = \hat{\eta}^{1,k}_0 + \hat{\eta}^{2,k}_0 = \overline{\eta}^{1,k}_0 + \overline{\eta}^{2,k}_0.$$

Thus it follows that $\hat{\chi}_k(0) = \overline{\chi}_k(0)$. By (3.2) and the local boundedness of the hazard rate functions $\{h^r_{1,k}, h^r_{2,k}, \ k \in \mathcal{K}\}$ assumed in Assumption 3.1, we have that for each $t \in [0,\infty)$ and $j = 1, 2$,

$$M^{j,r,k}_t \doteq \sup_{0\le u\le \overline{\chi}_k(0)\vee\hat{\chi}_k(0)+t} h^r_{j,k}(u) = \sup_{0\le u\le \overline{\chi}_k(0)+t} h^r_{j,k}(u) < \infty.$$

For each $k \in \mathcal{K}$, since $(\overline{X}^1_k, \overline{X}^2_k, \overline{\nu}^{1,k}, \overline{\nu}^{2,k}, \overline{\eta}^{1,k}, \overline{\eta}^{2,k})$ is a solution to the fluid model equations associated with the input data $(\overline{E}_k, \overline{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$ in Definition 2.1 of [14] and $(\hat{X}^1_k, \hat{X}^2_k, \hat{\nu}^{1,k}, \hat{\nu}^{2,k}, \hat{\eta}^{1,k}, \hat{\eta}^{2,k})$ is a solution to the fluid model equations associated with the input data $(\overline{E}_k, \hat{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$ in Definition 2.1 of [14], then the above display and Assumption 3.3 imply that Proposition 4.2 of [14] holds for the two solutions $(\overline{X}^1_k, \overline{X}^2_k, \overline{\nu}^{1,k}, \overline{\nu}^{2,k}, \overline{\eta}^{1,k}, \overline{\eta}^{2,k})$ and $(\hat{X}^1_k, \hat{X}^2_k, \hat{\nu}^{1,k}, \hat{\nu}^{2,k}, \hat{\eta}^{1,k}, \hat{\eta}^{2,k})$. Then for each $T > 0$, there exists a positive constant (can be chosen to be independent of $k \in \mathcal{K}$) $C_T < \infty$ such that

$$\sup_{t\in[0,T]} \left| (\overline{X}^1_k(t) + \overline{X}^2_k(t)) - (\hat{X}^1_k(t) + \hat{X}^2_k(t)) \right| \le C_T \sup_{w\in[0,T]} \left| \overline{I}_k(w) - \hat{I}_k(w) \right|. \tag{3.3}$$

It follows from (2.13) and (2.14) that for each $k \in \mathcal{K}$, $j = 1, 2$ and $t \ge 0$,

$$\left| \overline{Q}^j_k(t) - \hat{Q}^j_k(t) \right| = \left| F^{\overline{\eta}^{j,k}_t} \left((F^{\overline{\eta}^k_t})^{-1} (\overline{Q}^1_k(t) + \overline{Q}^2_k(t)) \right) - F^{\hat{\eta}^{j,k}_t} \left((F^{\hat{\eta}^k_t})^{-1} (\hat{Q}^1_k(t) + \hat{Q}^2_k(t)) \right) \right|.$$

It follows from Lemma 2.4 of [14], (3.29) of [14] and (4.79) and (4.86) of [14] that for each $T > 0$, $k \in \mathcal{K}$ and $j = 1, 2$, there exists a positive constant (can be chosen to be independent of $k \in \mathcal{K}$ and still denoted as C_T) $C_T < \infty$ such that

$$\sup_{t\in[0,T]} \left| \overline{Q}^j_k(t) - \hat{Q}^j_k(t) \right| \le C_T \sup_{w\in[0,T]} \left| \overline{I}_k(w) - \hat{I}_k(w) \right| \tag{3.4}$$

and

$$\sup_{t\in[0,T]} \left| \overline{R}^j_k(t) - \hat{R}^j_k(t) \right| \le C_T \sup_{w\in[0,T]} \left| \overline{I}_k(w) - \hat{I}_k(w) \right|. \tag{3.5}$$

By (2.21) and (2.22), we have that for each $T > 0$, $k \in \mathcal{K}$,

$$\begin{aligned} \sup_{t\in[0,T]} \left| \overline{L}^1_k(t) - \hat{L}^1_k(t) \right| &\le \sup_{t\in[0,T]} \left| \overline{Q}^1_k(t) - \hat{Q}^1_k(t)) \right| + \sup_{t\in[0,T]} \left| \overline{R}^1_k(t) - \hat{R}^1_k(t) \right| \\ &\le 2C_T \sup_{w\in[0,T]} \left| \overline{I}_k(w) - \hat{I}_k(w) \right| \end{aligned} \tag{3.6}$$

and

$$\begin{aligned}\sup_{t\in[0,T]}\left|\overline{L}_k^2(t)-\hat{L}_k^2(t))\right| &\le \sup_{t\in[0,T]}\left|\overline{Q}_k^2(t)-\hat{Q}_k^2(t))\right|+\sup_{t\in[0,T]}\left|\overline{R}_k^2(t)-\hat{R}_k^2(t)\right|+\sup_{w\in[0,T]}\left|\overline{I}_k(w)-\hat{I}_k(w)\right| \\ &\le 3C_T\sup_{w\in[0,T]}\left|\overline{I}_k(w)-\hat{I}_k(w)\right|. \end{aligned} \tag{3.7}$$

From (2.12), we see that for each $k\in\mathcal{K}$, $T>0$ and $t\in[0,T]$,

$$\left|\overline{I}_k(t)-\hat{I}_k(t)\right|\le\sum_{l\in\mathcal{K}}P_{lk}\left(\left|\int_0^t\left(\langle h_{1,l}^s,\overline{\nu}_u^{1,l}\rangle-\langle h_{1,l}^s,\hat{\nu}_u^{1,l}\rangle\right)du\right|+\left|\int_0^t\left(\langle h_{2,l}^s,\overline{\nu}_u^{2,l}\rangle-\langle h_{2,l}^s,\hat{\nu}_u^{2,l}\rangle\right)du\right|\right).$$

It follows from (2.6), (2.8), and an application of integration by parts that for each $l\in\mathcal{K}$, $j=1,2$, $T>0$ and $t\in[0,T]$,

$$\begin{aligned}\int_0^t\langle h_{j,l}^s,\overline{\nu}_u^{j,l}\rangle du &= \int_{[0,H_{j,l}^s)}\frac{G_{j,l}^s(x+t)-G_{j,l}^s(x)}{1-G_{j,l}^s(x)}\,\overline{\nu}_0^{j,l}(dx)+\int_0^t\int_0^s g_{j,l}^s(s-u)\,d\overline{L}_l^j(u)ds \\ &= \int_{[0,H_{j,l}^s)}\frac{G_{j,l}^s(x+t)-G_{j,l}^s(x)}{1-G_{j,l}^s(x)}\,\overline{\nu}_0^{j,l}(dx)+\int_0^t G_{j,l}^s(t-u)\,d\overline{L}_l^j(u) \\ &= \int_{[0,H_{j,l}^s)}\frac{G_{j,l}^s(x+t)-G_{j,l}^s(x)}{1-G_{j,l}^s(x)}\,\overline{\nu}_0^{j,l}(dx)+\int_0^t\overline{L}_l^j(u)g_{j,l}^s(t-u)\,du\end{aligned}$$

and

$$\int_0^t\langle h_{j,l}^s,\hat{\nu}_u^{j,l}\rangle du = \int_{[0,H_{j,l}^s)}\frac{G_{j,l}^s(x+t)-G_{j,l}^s(x)}{1-G_{j,l}^s(x)}\,\hat{\nu}_0^{j,l}(dx)+\int_0^t\hat{L}_l^j(u)g_{j,l}^s(t-u)\,du.$$

Note that $\hat{\nu}_0^{j,l}=\overline{\nu}_0^{j,l}$, it then follows that for each $l\in\mathcal{K}$, $j=1,2$, $T>0$ and $t\in[0,T]$,

$$\begin{aligned}\left|\int_0^t\left(\langle h_{j,l}^s,\overline{\nu}_u^{j,l}\rangle-\langle h_{j,l}^s,\hat{\nu}_u^{j,l}\rangle\right)du\right| &= \left|\int_0^t\overline{L}_l^j(u)g_{j,l}^s(t-u)\,du-\int_0^t\hat{L}_l^j(u)g_{j,l}^s(t-u)\,du\right| \\ &\le \int_0^t\left|\overline{L}_l^j(u)-\hat{L}_l^j(u)\right|g_{j,l}^s(t-u)\,du \\ &\le \sup_{u\in[0,T]}\left|\overline{L}_l^j(u)-\hat{L}_l^j(u)\right|G_{j,l}^s(t).\end{aligned}$$

Then we yield from (3.6) and (3.7) that for each $k\in\mathcal{K}$, $T>0$ and $t\in[0,T]$,

$$\begin{aligned}&\left|\overline{I}_k(t)-\hat{I}_k(t)\right| \\ \le&\sum_{l\in\mathcal{K}}P_{lk}\left(\sup_{u\in[0,T]}\left|\overline{L}_l^1(u)-\hat{L}_l^1(u)\right|G_{1,l}^s(t)+\sup_{u\in[0,T]}\left|\overline{L}_l^2(u)-\hat{L}_l^2(u)\right|G_{2,l}^s(t)\right) \\ \le&\sum_{l\in\mathcal{K}}P_{lk}\left(G_{1,l}^s(t)+G_{2,l}^s(t)\right)\left(\sup_{u\in[0,T]}\left|\overline{L}_l^2(u)-\hat{L}_l^2(u)\right|+\sup_{u\in[0,T]}\left|\overline{L}_l^2(u)-\hat{L}_l^2(u)\right|\right) \\ \le&\sum_{l\in\mathcal{K}}P_{lk}\left(G_{1,l}^s(t)+G_{2,l}^s(t)\right)5C_T\sup_{u\in[0,T]}\left|\overline{I}_k(u)-\hat{I}_k(u)\right|.\end{aligned} \tag{3.8}$$

Now choose $T>0$ such that

$$M_T\doteq\max_{k\in\mathcal{K}}\sum_{l\in\mathcal{K}}P_{lk}\left(G_{1,l}^s(T)+G_{2,l}^s(T)\right)5C_T<1.$$

Then we have that for each $k \in \mathcal{K}$,

$$\sup_{t\in[0,T]} \left|\overline{I}_k(t) - \hat{I}_k(t)\right| \le M_T \sup_{u\in[0,T]} \left|\overline{I}_k(u) - \hat{I}_k(u)\right|, \tag{3.9}$$

which in turn implies that $\overline{I}_k = \hat{I}_k$ on $[0,T]$ for each $k \in \mathcal{K}$. Then $(\overline{X}^1_k, \overline{X}^2_k, \overline{\nu}^{1,k}, \overline{\nu}^{2,k}, \overline{\eta}^{1,k}, \overline{\eta}^{2,k})$ and $(\hat{X}^1_k, \hat{X}^2_k, \hat{\nu}^{1,k}, \hat{\nu}^{2,k}, \hat{\eta}^{1,k}, \hat{\eta}^{2,k})$ restricted on $[0,T]$ are two solutions to the fluid model equations associated with the input data $(\overline{E}_k, \overline{I}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0)$ in Definition 2.1 of [14]. It follows from Theorem 3.6 of [14] that

$$(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2) = (\hat{X}^1, \hat{X}^2, \hat{\nu}^1, \hat{\nu}^2, \hat{\eta}^1, \hat{\eta}^2) \text{ on } [0,T].$$

By using Lemma 2.5 and applying a simple contradiction argument, it is clear that the maximal interval on which $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2) = (\hat{X}^1, \hat{X}^2, \hat{\nu}^1, \hat{\nu}^2, \hat{\eta}^1, \hat{\eta}^2)$ has to be $[0,\infty)$. This completes the proof of the uniqueness of solutions to the fluid model equations associated with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$.

We next show the existence of solutions to the fluid model equations associated with the input data $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0), \overline{\nu}^1_0, \overline{\nu}^2_0, \overline{\eta}^1_0, \overline{\eta}^2_0) \in \mathcal{S}_0$. For this, we shall construct the following sequence of processes $\{(\overline{X}^{n,1}, \overline{X}^{n,2}, \overline{\nu}^{n,1}, \overline{\nu}^{n,2}, \overline{\eta}^{n,1}, \overline{\eta}^{n,2}),\ n \ge 1\}$ recursively. For $n = 1$, we define $(\overline{X}^{1,1}, \overline{X}^{1,2}, \overline{\nu}^{1,1}, \overline{\nu}^{1,2}, \overline{\eta}^{1,1}, \overline{\eta}^{1,2})$ as follows. For each $k \in \mathcal{K}$, let

$$(\overline{X}^{1,1}_k, \overline{X}^{1,2}_k, \overline{\nu}^{1,1,k}, \overline{\nu}^{1,2,k}, \overline{\eta}^{1,1,k}, \overline{\eta}^{1,2,k})$$

be the solution to the fluid model equations in Definition 2.1 of [14] associated with the input data $(\overline{E}_k, \mathbf{0}, \overline{X}^1_k(0),\ \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0) \in \mathcal{S}_0$. For $n \ge 2$, define $(\overline{X}^{n,1}, \overline{X}^{n,2}, \overline{\nu}^{n,1}, \overline{\nu}^{n,2}, \overline{\eta}^{n,1}, \overline{\eta}^{n,2})$ such that for each $k \in \mathcal{K}$,

$$(\overline{X}^{n,1}_k, \overline{X}^{n,2}_k, \overline{\nu}^{n,1,k}, \overline{\nu}^{n,2,k}, \overline{\eta}^{n,1,k}, \overline{\eta}^{n,2,k})$$

is the solution to the fluid model equations in Definition 2.1 of [14] associated with the input data $(\overline{E}_k, \overline{I}^{n-1}_k, \overline{X}^1_k(0), \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0) \in \mathcal{S}_0$, where for each $t \in [0,\infty)$,

$$\overline{I}^{n-1}_k(t) = \sum_{l\in\mathcal{K}} P_{lk} \int_0^t \left(\langle h^s_{1,l}, \overline{\nu}^{n-1,1,l}_u \rangle + \langle h^s_{2,l}, \overline{\nu}^{n-1,2,l}_u \rangle \right) du. \tag{3.10}$$

Then we can apply the same argument for (3.9) with $(\overline{X}^{n+2,1}, \overline{X}^{n+2,2}, \overline{\nu}^{n+2,1}, \overline{\nu}^{n+2,2}, \overline{\eta}^{n+2,1}, \overline{\eta}^{n+2,2})$ and $(\overline{X}^{n+1,1}, \overline{X}^{n+1,2}, \overline{\nu}^{n+1,1}, \overline{\nu}^{n+1,2}, \overline{\eta}^{n+1,1}, \overline{\eta}^{n+1,2})$, $n \ge 1$, in place of $(\overline{X}^1_k, \overline{X}^2_k, \overline{\nu}^{1,k}, \overline{\nu}^{2,k}, \overline{\eta}^{1,k}, \overline{\eta}^{2,k})$ and $(\hat{X}^1_k, \hat{X}^2_k, \hat{\nu}^{1,k}, \hat{\nu}^{2,k}, \hat{\eta}^{1,k}, \hat{\eta}^{2,k})$, respectively, to get that there exists $T > 0$ such that for each $n \ge 1$,

$$\max_{k\in\mathcal{K}} \sup_{t\in[0,T]} \left|\overline{I}^{n+2}_k(t) - \overline{I}^{n+1}_k(t)\right| \le M_T \left(\max_{k\in\mathcal{K}} \sup_{w\in[0,T]} \left|\overline{I}^{n+1}_k(w) - \overline{I}^{n}_k(w)\right| \right). \tag{3.11}$$

This implies that the sequence $\{\overline{I}^n, n \ge 1\}$ converges uniformly on $[0,T]$. Let $\overline{I}^*$ denote the limit of $\{\overline{I}^n, n \ge 1\}$ on $[0,T]$. Now, for each $k \in \mathcal{K}$, let

$$(\overline{X}^{*,1}_k, \overline{X}^{*,2}_k, \overline{\nu}^{*,1,k}, \overline{\nu}^{*,2,k}, \overline{\eta}^{*,1,k}, \overline{\eta}^{*,2,k})$$

be the solution to the fluid model equations in Definition 2.1 of [14] associated with the input data $(\overline{E}_k, \overline{I}^*_k, \overline{X}^1_k(0),\ \overline{X}^2_k(0), \overline{\nu}^{1,k}_0, \overline{\nu}^{2,k}_0, \overline{\eta}^{1,k}_0, \overline{\eta}^{2,k}_0) \in \mathcal{S}_0$. To show that $(\overline{X}^{*,1}, \overline{X}^{*,2}, \overline{\nu}^{*,1}, \overline{\nu}^{*,2}, \overline{\eta}^{*,1}, \overline{\eta}^{*,2})$ is a solution to the fluid model equations, we just need to show that (2.12) holds. Note that for each $k \in \mathcal{K}$, by Theorem 4.1 of [14], as $n \to \infty$, $(\overline{X}^{n,1}_k, \overline{X}^{n,2}_k, \overline{Q}^{n,1}_k, \overline{Q}^{n,2}_k, \overline{R}^{n,1}_k, \overline{R}^{n,2}_k)$ converges uniformly on $[0,T]$ to $(\overline{X}^{*,1}_k, \overline{X}^{*,2}_k, \overline{Q}^{*,1}_k, \overline{Q}^{*,2}_k, \overline{R}^{*,1}_k, \overline{R}^{*,2}_k)$. Since for each $j \in \mathcal{K}$,

$$\int_0^t \langle h^s_{1,l}, \overline{\nu}^{n,1,l}_u \rangle\, du \quad = \quad \int_{[0,H^s_{1,l})} \frac{G^s_{1,l}(x+t) - G^s_{1,l}(x)}{1 - G^s_{1,l}(x)}\, \overline{\nu}^{1,l}_0(dx)$$

$$
\begin{aligned}
&\quad + \int_0^t \left(\overline{Q}_l^1(0) + \overline{E}_l(u) - \overline{Q}_l^{n,1}(u) - \overline{R}_l^{n,1}(u) \right) g_{1,l}^s(t-u)du \\
&\to \int_{[0,H_{1,l}^s)} \frac{G_{1,l}^s(x+t) - G_{1,l}^s(x)}{1 - G_{1,l}^s(x)} \, \overline{\nu}_0^{1,l}(dx) \\
&\quad + \int_0^t \left(\overline{Q}_l^1(0) + \overline{E}_l(u) - \overline{Q}_l^{*,1}(u) - \overline{R}_l^{*,1}(u) \right) g_{1,l}^s(t-u)du \\
&= \int_0^t \langle h_{1,l}^s, \overline{\nu}_u^{*,1,l} \rangle \, du \quad \text{uniformly for all } t \in [0,T],
\end{aligned}
$$

and

$$
\begin{aligned}
\int_0^t \langle h_{2,l}^s, \overline{\nu}_u^{n,2,l} \rangle \, du &= \int_{[0,H_{2,l}^s)} \frac{G_{2,l}^s(x+t) - G_{2,l}^s(x)}{1 - G_{2,l}^s(x)} \, \overline{\nu}_0^{2,l}(dx) \\
&\quad + \int_0^t \left(\overline{Q}_l^2(0) + \overline{I}_l^n(u) - \overline{Q}_l^{n,2}(u) - \overline{R}_l^{n,2}(u) \right) g_{2,l}^s(t-u)du \\
&\to \int_{[0,H_{2,l}^s)} \frac{G_{2,l}^s(x+t) - G_{2,l}^s(x)}{1 - G_{2,l}^s(x)} \, \overline{\nu}_0^{2,l}(dx) \\
&\quad + \int_0^t \left(\overline{Q}_l^2(0) + \overline{I}_l^*(u) - \overline{Q}_l^{*,2}(u) - \overline{R}_l^{*,2}(u) \right) g_{2,l}^s(t-u)du \\
&= \int_0^t \langle h_{2,l}^s, \overline{\nu}_u^{*,2,l} \rangle \, du \quad \text{uniformly for all } t \in [0,T].
\end{aligned}
$$

Then by taking the limits on both sides of the display (3.10), we have that for each $k \in \mathcal{K}$,

$$
\overline{I}_k^*(t) = \sum_{l \in \mathcal{K}} P_{lk} \int_0^t \left(\langle h_{1,l}^s, \overline{\nu}_u^{*,1,l} \rangle + \langle h_{2,l}^s, \overline{\nu}_u^{*,2,l} \rangle \right) du.
$$

This shows that (2.12 holds and $(\overline{X}^{*,1}, \overline{X}^{*,2}, \overline{\nu}^{*,1}, \overline{\nu}^{*,2}, \overline{\eta}^{*,1}, \overline{\eta}^{*,2})$ is a solution to the fluid model equations. □

4. Invariant States

Given a positive constant vector $\lambda = (\lambda_k : k \in \mathcal{K})$, a state $(x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2)$ such that $(e_\lambda, x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2) \in \mathcal{S}_0$ and η_0^1, η_0^2 are continuous on $\mathbb{R}_+$ is said to be an *invariant state* for the fluid model equations in Definition 2.1 if there is a solution $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ to the fluid model equations associated with the initial data $(e_\lambda, x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2) \in \mathcal{S}_0$ satisfies $(\overline{X}^1(t), \overline{X}^2(t), \overline{\nu}_t^1, \overline{\nu}_t^2, \overline{\eta}_t^1, \overline{\eta}_t^2) = (x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2)$ for all $t \geq 0$, where $e_\lambda(t) = \lambda t$ for each $t \geq 0$.

Let $\nu_*^{j,k}$ and $\eta_*^{j,k}$, $k \in \mathcal{K}$, $j = 1, 2$, be the non-negative measures defined as follows:

$$
\nu_*^{j,k}[0,x) = \int_0^x (1 - G_{j,k}^s(y))dy, \quad x \in [0, H_{j,k}^s), \tag{4.1}
$$

$$
\eta_*^{j,k}[0,x) = \int_0^x (1 - G_{j,k}^r(y))dy, \quad x \in [0, H_{j,k}^r). \tag{4.2}
$$

Note that (2.1) implies that for each $K \in \mathcal{K}$, $\nu_*^{1,k}$ and $\nu_*^{2,k}$ are actually finite measures. Define the vector $\overline{\lambda}$ by

$$
\overline{\lambda} = (I - P')^{-1}\lambda = H\lambda. \tag{4.3}
$$

Namely, the kth entry $\overline{\lambda}_k$ is the effective/overall arrival rate of customers of class k.

Define

$$
\mathcal{K}^\dagger := \{k \in \mathcal{K} : \ \lambda_k m_{1,k}^s + (\overline{\lambda}_k - \lambda_k) m_{2,k}^s \geq s_k\},
$$

the set of potentially critically loaded or overloaded service stations in the absence of impatience. Note that $k \notin \mathcal{K}^\dagger$ if $s_k = \infty$, which says that any service station with infinite servers cannot be potentially critically loaded or overloaded.

Let w and χ be K-dimensional non-negative vectors satisfying the following equation and vector inequality:

$$w = P'G^1(\chi)\lambda + P'G^2(\chi)w, \tag{4.4}$$

$$\mathrm{diag}(m_1^s)G^1(\chi)\lambda + \mathrm{diag}(m_2^s)G^2(\chi)w \le s, \tag{4.5}$$

where $G^1(\chi)$ is the $K \times K$ diagonal matrix with its kth diagonal entry $1 - G^r_{1,k}(\chi_k)$ and $G^2(\chi)$ is the $K \times K$ diagonal matrix with its kth diagonal entry $1 - G^r_{2,k}(\chi_k)$. Note that

$$\begin{aligned}(I - G^2(\chi)P')^{-1} &= 1 + G^2(\chi)P' + (G^2(\chi)P')^2 + (G^2(\chi)P')^3 + \cdots \\ &\le I + P' + (P')^2 + (P')^3 + \cdots = (I - P')^{-1}\end{aligned}$$

and by (4.4),

$$w = (I - G^2(\chi)P')^{-1}P'G^1(\chi)\lambda = P'(I - G^2(\chi)P')^{-1}G^1(\chi)\lambda \le P'(I - P')^{-1}\lambda.$$

It follows that for any K-dimensional non-negative vector χ,

$$\begin{aligned}& \mathrm{diag}(m_1^s)G^1(\chi)\lambda + \mathrm{diag}(m_2^s)G^2(\chi)w \\ \le\ & \mathrm{diag}(m_1^s)\lambda + \mathrm{diag}(m_2^s)P'(I - P')^{-1}\lambda \\ =\ & \mathrm{diag}(m_1^s)\lambda + \mathrm{diag}(m_2^s)\left((I - P')^{-1} - I\right)\lambda = \mathrm{diag}(m_1^s)\lambda + \mathrm{diag}(m_2^s)\left(\overline{\lambda} - \lambda\right).\end{aligned}$$

This implies that for all $k \notin \mathcal{K}^\dagger$, the kth entry of the vector $\mathrm{diag}(m_1^s)G^1(\chi)\lambda + \mathrm{diag}(m_2^s)G^2(\chi)w$ is strictly less than s_k. Let $\mathcal{Z}$ be the set defined by

$$\mathcal{Z} \doteq \left\{ (w,\chi) \in \mathbb{R}_+^K \times \mathbb{R}_+^K : \begin{array}{c} \lambda_k(1 - G^r_{1,k}(\chi_k))m^s_{1,k} + w_k(1 - G^r_{2,k}(\chi_k))m^s_{2,k} = s_k \text{ for each } k \in \mathcal{J} \\ \text{and } \lambda_k m^s_{1,k} + w_k m^s_{2,k} < s_k \text{ and } \chi_k = 0 \text{ for each } k \in \mathcal{K}\setminus\mathcal{J}, \\ \text{where } \mathcal{J} \subseteq \mathcal{K}^\dagger \text{ and } (w,\chi) \text{ satisfies (4.4 and (4.5)} \end{array} \right\}, \tag{4.6}$$

and for each $(w,\chi) \in \mathcal{Z}$,

$$\mathcal{X}_{(w,\chi)} = \left\{ (x^1,x^2) \in \mathbb{R}_+^K \times \mathbb{R}_+^K : \begin{array}{l} x^1_k = \lambda_k(1 - G^r_{1,k}(\chi_k(0)))m^s_{1,k} + \lambda_k \int_0^{\chi_k(0)} (1 - G^r_{1,k}(x))dx, \\ x^2_k = w_k(1 - G^r_{2,k}(\chi_k(0)))m^s_{2,k} + w_k \int_0^{\chi_k(0)} (1 - G^r_{2,k}(x))dx \\ \text{for each } k \in \mathcal{K} \end{array} \right\}. \tag{4.7}$$

Let $\mathcal{I}_\lambda$ be the set of states defined by

$$\mathcal{I}_\lambda = \left\{ \begin{array}{c} (x^1,x^2,G^1(\chi)\mathrm{diag}(\lambda)\nu^1_*, G^2(\chi)\mathrm{diag}(w)\nu^2_*, \mathrm{diag}(\lambda)\eta^1_*, \mathrm{diag}(w)\eta^2_*) : \\ (w,\chi) \in \mathcal{Z},\ x = (x^1,x^2) \in \mathcal{X}_{(w,\chi)} \end{array} \right\}. \tag{4.8}$$

It is clear that, if $\mathcal{K}^\dagger = \emptyset$, then $\mathcal{Z}$ contains only one element $(\overline{\lambda} - \lambda, \mathbf{0})$. On the other hand, if $\mathcal{K}^\dagger \neq \emptyset$, the set $\mathcal{J}$ in (4.6) should be non-empty. Note that for each $k \in \mathcal{K}^\dagger$, the presence of customers' impatience may reduce the actual arrival rate to service station k from internal routing. Then the actual arrival rate to service station k may be less than the effective arrival rate $\bar{\lambda}_k$. As a result, the set $\mathcal{J}$ in (4.6) could be a strict subset of $\mathcal{K}^\dagger$. In general, the set $\mathcal{Z}$ may contain several elements depending on the choices of $\mathcal{J}$.

Remark 4.1. *When customers in the system have infinite patience, that is, $G^r_{j,k}(x) = 0$ for all $x \in [0,\infty)$, $k \in \mathcal{K}$ and $j = 1,2$, and $\mathcal{K}^\dagger = \mathcal{K} = \{k \in \mathcal{K} : \ \lambda_k m^s_{1,k} + (\bar{\lambda}_k - \lambda_k)m^s_{2,k} > s_k\}$, that is, all service stations are overloaded, the invariant state $\mathcal{I}_\lambda$ is actually an empty set. This is consistent with the fact that the overloaded system in the absence of impatience is not stable. In fact, suppose that $\mathcal{Z} \neq \emptyset$. Then, for each $(w,\chi) \in \mathcal{Z}$, (w,χ) satisfies (4.4). Note that, in this*

case, (4.4) is reduced to the equation $w = P'\lambda + P'w$ and this implies that $w = \overline{\lambda} - \lambda$. Since $\lambda_k m^s_{1,k} + (\bar{\lambda}_k - \lambda_k) m^s_{2,k} > s_k$ for each $k \in \mathcal{K}$, then $\mathcal{J}$ in (4.6) does not exist. This is a contradiction to the fact that $(w, \chi) \in \mathcal{Z}$. Thus, when all stations are overloaded, the set $\mathcal{Z}$ in (4.6) is an empty set, and so is $\mathcal{I}_\lambda$.

Remark 4.2. *If the system has only one critically loaded or overloaded service station, that is, $\mathcal{K}^\dagger$ has only one element, or every service station in the system is underloaded, that is, $\mathcal{K}^\dagger = \emptyset$, and if, in addition, the patience time distributions $G^r_{1,k}$ and $G^r_{2,k}$, $k \in \mathcal{K}$, are strictly increasing, then $\mathcal{Z}$ has only one element and the system has only one invariant state.*

Remark 4.3. *If the system has feed-forward routing, that is, the routing matrix P has the property that $P_{ij} = 0$ for each $i \geq j$ (then P' is a strictly lower triangular matrix) and if the patience time distributions are all strictly increasing, then the set $\mathcal{Z}$ has a unique element and hence the system has a unique invariant state. In fact, if $\mathcal{K}^\dagger \neq \emptyset$, let $\mathcal{K}^\dagger$ be the increasingly ordered set $\{k_1, k_2, \cdots, k_n\}$ for some $n \leq K$ and $\mathcal{J}$ be a non-empty subset of $\mathcal{K}^\dagger$, which is associated with an invariant state. Note that by the definition of $\mathcal{Z}$, $\chi_i = 0$ for all $i < k_1$. Since P' is strictly lower triangular, we can write (4.4) in terms of block matrices:*

$$\begin{bmatrix} \tilde{w} \\ \hat{w} \end{bmatrix} = \begin{bmatrix} A & 0 \\ B & C \end{bmatrix} \begin{bmatrix} \tilde{G}^1(\tilde{\chi}) & 0 \\ 0 & \hat{G}^1(\hat{\chi}) \end{bmatrix} \begin{bmatrix} \tilde{\lambda} \\ \hat{\lambda} \end{bmatrix} + \begin{bmatrix} A & 0 \\ B & C \end{bmatrix} \begin{bmatrix} \tilde{G}^2(\tilde{\chi}) & 0 \\ 0 & \hat{G}^2(\hat{\chi}) \end{bmatrix} \begin{bmatrix} \tilde{w} \\ \hat{w} \end{bmatrix},$$

where

$$P' = \begin{bmatrix} A & 0 \\ B & C \end{bmatrix}.$$

By the block matrix multiplications, we have that

$$\tilde{w} = A\tilde{G}^1(\tilde{\chi})\tilde{\lambda} + A\tilde{G}^2(\tilde{\chi})\tilde{w}. \tag{4.9}$$

Since A is strictly lower triangular and $\chi_i = 0$ for all $i < k_1$ (here the value of χ_{k_1} in (4.9) is not relevant since A is strictly lower triangular), (4.9) reduces to

$$\tilde{w} = A\tilde{\lambda} + A\tilde{w}, \text{ that is, } \tilde{w} = (I - A)^{-1}A\tilde{\lambda} = (I - A)^{-1}\tilde{\lambda} - \tilde{\lambda}.$$

It follows from (4.3) that

$$\overline{\lambda} = \begin{bmatrix} I - A & 0 \\ -B & I - C \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\lambda} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} (I - A)^{-1} & 0 \\ (I - C)^{-1}B(I - A)^{-1} & (I - C)^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\lambda} \\ \hat{\lambda} \end{bmatrix}.$$

Combining the above two displays, we can see that when $\overline{w}$ in the above block representation of (4.4) represents the first k_1 entries of w, $w_i = \overline{\lambda}_i - \lambda_i$ for all $i \leq k_1$. It follows from (4.5) that

$$\begin{aligned} & m^s_{1,k_1}(1 - G^r_{1,k_1}(\chi_{k_1}))\lambda_{k_1} + m^s_{2,k_1}(1 - G^r_{2,k_1}(\chi_{k_1}))w_{k_1} \\ = \; & m^s_{1,k_1}(1 - G^r_{1,k_1}(\chi_{k_1}))\lambda_{k_1} + m^s_{2,k_1}(1 - G^r_{2,k_1}(\chi_{k_1}))(\overline{\lambda}_{k_1} - \lambda_{k_1}) \leq s_{k_1}. \end{aligned}$$

Note that $k_1 \in \mathcal{K}^\dagger$. Then $k_1 \in \mathcal{J}$ since, otherwise, by the definition of $\mathcal{Z}$, $\chi_{k_1} = 0$ and $\lambda_{k_1} m^s_{1,k_1} + (\overline{\lambda}_{k_1} - \lambda_{k_1}) m^s_{2,k_1} < s_{k_1}$. This contradicts the fact that $k_1 \in \mathcal{K}^\dagger$. Since $k_1 \in \mathcal{J}$, then there exists a unique $\chi_{k_1} \geq 0$ such that

$$m^s_{1,k_1}(1 - G^r_{1,k_1}(\chi_{k_1}))\lambda_{k_1} + m^s_{2,k_1}(1 - G^r_{2,k_1}(\chi_{k_1}))(\overline{\lambda}_{k_1} - \lambda_{k_1}) = s_{k_1}.$$

For any service station i such that $k_1 < i < k_2$, it will remain underloaded since the input rate from service station k_1 to service station i is less than or equal to the effective input rate due to abandonment at service station k_1, then for all $k_1 < i < k_2$, $\chi_i = 0$ and $w_i \leq \overline{\lambda}_i - \lambda_i$ is uniquely determined by $\lambda_j, s_j, m^s_{1,j}, m^s_{2,j}$ for $j \leq i$ and P (here w_i also depends on χ_{k_1}, which is uniquely determined by those parameters as well). For service station k_2, we know that the effective system utilization on service station k_2 is $\lambda_{k_2} m^s_{1,k_2} + (\overline{\lambda}_{k_2} - \lambda_{k_2}) m^s_{2,k_2} \geq s_{k_2}$ and the actual system utilization on service station k_2 is $\lambda_{k_2} m^s_{1,k_2} + w_{k_2} m^s_{2,k_2}$, where w_{k_2} is uniquely determined

by $\lambda_j, s_j, m^s_{1,j}, m^s_{2,j}$ for $j \le k_2$ and P through (4.9) when $\overline{w}$ in (4.9) represents the first k_2 entries of w (here again the value of χ_{k_2} in (4.9) is not relevant since A is strictly lower triangular). So if $\lambda_{k_2} m^s_{1,k_2} + w_{k_2} m^s_{2,k_2} \ge s_{k_2}$, then $k_2 \in \mathcal{J}$ and there exists a unique $\chi_{k_2} \ge 0$ such that

$$m^s_{1,k_2}(1 - G^r_{1,k_2}(\chi_{k_2}))\lambda_{k_2} + m^s_{2,k_2}(1 - G^r_{2,k_2}(\chi_{k_2}))w_{k_2} = s_{k_2};$$

otherwise, $k_2 \notin \mathcal{J}$. Note that whether k_2 is in $\mathcal{J}$ or not depends only on $\lambda_j, s_j, m^s_{1,j}, m^s_{2,j}$ for $j \le k_2$ and P. By a similar argument for the rest of the stations in $\mathcal{K}^\dagger$, we can see that the choice of $\mathcal{J}$ and the values of w, χ are uniquely determined by $\lambda_j, s_j, m^s_{1,j}, m^s_{2,j}$ for $j \in \mathcal{K}$ and P. Thus, there is a unique element in $\mathcal{Z}$ and the system has a unique invariant state.

Theorem 4.4. (*Characterization of the Invariant States*) *Given the arrival rate vector $\lambda = (\lambda_k : k \in \mathcal{K})$, the set $\mathcal{I}_\lambda$ gives all invariant states associated with the fluid equations (2.5) – (2.19).*

Proof. Fix the arrival rate vector $\lambda = (\lambda_k : k \in \mathcal{K})$ and let $e_\lambda(t) = \lambda t$ for each $t \ge 0$. We break the argument into the following two claims.

Claim 1. The set of invariant states is a subset of $\mathcal{I}_\lambda$. Let $(x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2)$ be an invariant state and $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ be the solution to the fluid model equations associated with the initial data $(e_\lambda, x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2)$ that satisfies $(\overline{X}^1_t, \overline{X}^2_t, \overline{\nu}^1_t, \overline{\nu}^2_t, \overline{\eta}^1_t, \overline{\eta}^2_t) = (x^1, x^2, \nu^1, \nu^2, \eta^1, \eta^2)$ for all $t \ge 0$. We will show that
(i) $\eta^{1,k} = \lambda_k \eta^{1,k}_*$ for each $k \in \mathcal{K}$,
(ii) $\nu^{1,k} = \lambda_k(1-G^r_{1,k}(\chi_k)\nu^{1,k}_*$, $\nu^{2,k} = w_k(1-G^r_{2,k}(\chi_k)\nu^{2,k}_*$, $\eta^{2,k} = w_k \eta^{2,k}_*$, $x^1_k = \lambda_k(1-G^r_{1,k}(\chi_k))m^s_{1,k} + \lambda_k \int_0^{\chi_k}(1-G^r_{1,k}(x))dx$ and $x^2_k = w_k(1-G^r_{2,k}(\chi_k))m^s_{2,k} + w_k \int_0^{\chi_k}(1-G^r_{2,k}(x))dx$ for some $(w, \chi) \in \mathcal{Z}$ and $(x^1, x^2) \in \mathcal{X}_{(w,\chi)}$.

To establish (i), fix $k \in \mathcal{K}$. Since $\overline{\eta}^{1,k}_t = \eta^{1,k}$ for each $t \ge 0$, it follows from (2.10) that for every $f \in \mathcal{C}_c(\mathbb{R}_+)$,

$$\begin{aligned}\int_{[0,H^r_{1,k})} f(x)\,\eta^{1,k}(dx) &= \int_{[0,H^r_{1,k})} f(x+t)\frac{1-G^r_{1,k}(x+t)}{1-G^r_{1,k}(x)}\,\eta^{1,k}(dx) \\ &\quad + \lambda_k \int_0^t f(t-s)(1-G^r_{1,k}(t-s))\,ds.\end{aligned}$$

Letting $t \to \infty$ and using the fact that f has compact support, we obtain

$$\int_{[0,H^r_{1,k})} f(x)\,\eta^{1,k}(dx) = \lambda_k \int_{[0,H^r_{1,k})} f(s)(1-G^r_{1,k}(s))\,ds,$$

which implies that

$$\eta^{1,k}(dx) = \lambda_k(1-G^r_{1,k}(x))\,dx = \lambda_k \eta^{1,k}_*(dx). \tag{4.10}$$

This establishes (i).

Next, we focus on establishing (ii). For each $k \in \mathcal{K}$, let

$$\chi_k \doteq (F^{\eta^k})^{-1}((x^1_k + x^2_k - s_k)^+), \text{ where } \eta^k \doteq \eta^{1,k} + \eta^{2,k}.$$

Since $\bar{X}^1_k(t) = x^1_k$ and $\bar{X}^2_k(t) = x^2_k$ for each $t \ge 0$ and $k \in \mathcal{K}$, by (2.20), we have $\overline{Q}^1_k(t) + \overline{Q}^2_k(t) = (x^1_k + x^2_k - s_k)^+$ for each $t \ge 0$ and $k \in \mathcal{K}$. Since $\overline{\eta}^{j,k}_t = \eta^{j,k}$ for each $t \ge 0$, $k \in \mathcal{K}$ and $j = 1, 2$, we also have that $\overline{\eta}^k_t = \overline{\eta}^{1,k}_t + \overline{\eta}^{2,k}_t = \eta^{1,k} + \eta^{2,k} = \eta^k$ for each $t \ge 0$ and $k \in \mathcal{K}$. This implies, in particular, that for each $t \ge 0$ and $k \in \mathcal{K}$,

$$\overline{\chi}_k(t) = (F^{\overline{\eta}^k_t})^{-1}(\overline{Q}^1_k(t) + \overline{Q}^2_k(t)) = (F^{\eta^k})^{-1}((x^1_k + x^2_k - s_k)^+) = \chi_k. \tag{4.11}$$

Since $\overline{\chi}_k(\cdot)$ is constant by (4.11) and $\overline{\eta}_t^{j,k} = \eta^{j,k}$ for each $t \geq 0$, $k \in \mathcal{K}$ and $j = 1, 2$, then by (2.15) and (2.16), we have that $\overline{R}_k^1(t) = c_{1,k}t$, and $\overline{R}_k^2(t) = c_{2,k}t$ for each $t \geq 0$ and $k \in \mathcal{K}$, where by (4.10),

$$c_{1,k} \doteq \int_{[0,H_{1,k}^r)} \mathbb{1}_{[0,\chi_k]}(u) h_{1,k}^r(u) \eta^{1,k}(du) = \lambda_k G_{1,k}^r(\chi_k), \tag{4.12}$$

and

$$c_{2,k} \doteq \int_{[0,H_{2,k}^r)} \mathbb{1}_{[0,\chi_k]}(u) h_{2,k}^r(u) \eta^{2,k}(du). \tag{4.13}$$

For each $k \in \mathcal{K}$, let

$$w_k \doteq \sum_{l \in \mathcal{K}} P_{lk} \left(\langle h_{1,l}^s, \nu^{1,l} \rangle + \langle h_{2,l}^s, \nu^{2,l} \rangle \right). \tag{4.14}$$

Then it follows from (2.12) that for each $k \in \mathcal{K}$ and $t \geq 0$,

$$\overline{I}_k(t) = w_k t. \tag{4.15}$$

Thus, by (2.21), we obtain that for each $t \geq 0$ and $k \in \mathcal{K}$, $\bar{L}_k^1(t) = (\lambda_k - c_{1,k})t$ and $\bar{L}_k^2(t) = (w_k - c_{2,k})t$. Now for each $k \in \mathcal{K}$, by letting $\bar{\nu}_t^{1,k} = \nu^{1,k}$ for each $t \geq 0$ in (2.6) and $\bar{\nu}_t^{2,k} = \nu^{2,k}$ for each $t \geq 0$ in (2.8), we see that for every $f \in \mathcal{C}_c(\mathbb{R}_+)$,

$$\begin{aligned} \int_{[0,H_{1,k}^s)} f(x) \nu^{1,k}(dx) &= \int_{[0,H_{1,k}^s)} f(x+t) \frac{1 - G_{1,k}^s(x+t)}{1 - G_{1,k}^s(x)} \nu^{1,k}(dx) \\ &\quad + (\lambda_k - c_{1,k}) \int_0^t f(u)(1 - G_{1,k}^s(u))\, du, \end{aligned} \tag{4.16}$$

and

$$\begin{aligned} \int_{[0,H_{2,k}^s)} f(x) \nu^{2,k}(dx) &= \int_{[0,H_{2,k}^s)} f(x+t) \frac{1 - G_{2,k}^s(x+t)}{1 - G_{2,k}^s(x)} \nu^{2,k}(dx) \\ &\quad + (w_k - c_{2,k}) \int_0^t f(u)(1 - G_{2,k}^s(u))\, du. \end{aligned} \tag{4.17}$$

Again, letting $t \to \infty$ on both sides of the above two displays and using the fact that f has compact support, we yield that for each $k \in \mathcal{K}$,

$$\int_{[0,H_{1,k}^s)} f(x) \nu^{1,k}(dx) = (\lambda_k - c_{1,k}) \int_{[0,H_{1,k}^s)} f(u)(1 - G_{1,k}^s(u))\, du, \tag{4.18}$$

and

$$\int_{[0,H_{2,k}^s)} f(x) \nu^{2,k}(dx) = (w_k - c_{2,k}) \int_{[0,H_{2,k}^s)} f(u)(1 - G_{2,k}^s(u))\, du. \tag{4.19}$$

This implies that for each $k \in \mathcal{K}$,

$$\nu^{1,k}(dx) = (\lambda_k - c_{1,k})(1 - G_{1,k}^s(x))dx = (\lambda_k - c_{1,k})\nu_*^{1,k}(dx) = \lambda_k(1 - G_{1,k}^r(\overline{\chi}_k))\nu_*^{1,k}(dx) \tag{4.20}$$

and

$$\nu^{2,k}(dx) = (w_k - c_{2,k})(1 - G_{2,k}^s(x))dx = (w_k - c_{2,k})\nu_*^{2,k}(dx). \tag{4.21}$$

Since $\langle \mathbf{1}, \nu^{1,k} \rangle + \langle \mathbf{1}, \nu^{2,k} \rangle \leq s_k$ for each $k \in \mathcal{K}$, it follows from (2.1), (4.20) and (4.21) that for each $k \in \mathcal{K}$,

$$(\lambda_k - c_{1,k}) m_{1,k}^s + (w_k - c_{2,k}) m_{2,k}^s \leq s_k. \tag{4.22}$$

It follows from (4.20) and (4.21) that for each $l \in \mathcal{K}$, $\langle h_{1,l}^s, \nu^{1,l} \rangle + \langle h_{2,l}^s, \nu^{2,l} \rangle = \lambda_l - c_{1,l} + w_l - c_{2,l}$. Thus, plugging this into (4.14), we have that for each $k \in \mathcal{K}$,

$$w_k = \sum_{l \in \mathcal{K}} P_{lk} \left(\lambda_l - c_{1,l} + w_l - c_{2,l} \right). \tag{4.23}$$

Next, for each $k \in \mathcal{K}$, by letting $\overline{\eta}_t^{2,k} = \eta^{2,k}$ for each $t \geq 0$ in (2.11) and using (4.15), we see that for every $f \in \mathcal{C}_c(\mathbb{R}_+)$,

$$\int_{[0,H_{2,k}^r)} f(x)\eta^{2,k}(dx) = \int_{[0,H_{2,k}^r)} f(x+t)\frac{1-G_{2,k}^r(x+t)}{1-G_{2,k}^r(x)}\eta^{2,k}(dx) + w_k \int_0^t f(s)(1-G_{2,k}^r(s))\,ds,$$

and then letting $t \to \infty$ and using the fact that f has compact support, we obtain that

$$\int_{[0,H_{2,k}^r)} f(x)\eta^{2,k}(dx) = w_k \int_{[0,H_{2,k}^r)} f(s)(1-G_{2,k}^r(s))\,ds.$$

This implies that for each $k \in \mathcal{K}$,

$$\eta^{2,k}(dx) = w_k(1-G_{2,k}^r(x))dx = w_k\eta_*^{2,k}(dx). \tag{4.24}$$

Combining (4.24) with (4.13), we have that for each $k \in \mathcal{K}$,

$$c_{2,k} = w_k G_{2,k}^r(\chi_k). \tag{4.25}$$

It follows from (4.12), (4.25) and (4.23) that

$$w_k = \sum_{l\in\mathcal{K}} P_{lk}\left(\lambda_l - \lambda_l G_{1,l}^r(\chi_l) + w_l - w_l G_{2,l}^r(\chi_l)\right). \tag{4.26}$$

Let w be the K-dimensional non-negative vector with its kth entry w_k, $G^1(\chi)$ be the $K\times K$ diagonal matrix with its kth diagonal entry $1-G_{1,k}^r(\chi_k)$ and $G^2(\chi)$ be the $K\times K$ diagonal matrix with its kth diagonal entry $1-G_{2,k}^r(\chi_k)$. Then (4.26) can be rewritten as

$$w = P'G^1(\chi)\lambda + P'G^2(\chi)w. \tag{4.27}$$

Thus, we see that w and χ satisfy (4.4). Then it from (4.12), (4.25) and (4.22), w and χ also satisfy (4.5). Now, let

$$\mathcal{J} = \left\{k \in \mathcal{K}^\dagger :\ \lambda_k(1-G_{1,k}^r(\chi_k))m_{1,k}^s + w_k(1-G_{2,k}^r(\chi_k))m_{2,k}^s = s_k\right\}.$$

Then for each $k \in \mathcal{K}\setminus\mathcal{J}$, by (4.22), we have that $\langle \mathbf{1}, \nu^{1,k}\rangle + \langle \mathbf{1}, \nu^{2,k}\rangle < s_k$ and hence $(x_k^1 + x_k^2 - s_k)^+ = 0$ by (2.19) and then $\chi_k = (F^{\eta^k})^{-1}(0)$. Note that

$$\eta^k(dx) = \lambda_k(1-G_{1,k}^r(x))dx + w_k(1-G_{2,k}^r(x))dx.$$

It follows that $\chi_k = (F^{\eta^k})^{-1}(0) = 0$ and then, $\lambda_k m_{1,k}^s + w_k m_{2,k}^s < s_k$. Therefore, $(w,\chi) \in \mathcal{Z}$, which is defined in (4.6). It follows from (2.13) and (2.14) that

$$x_k^1 = \langle \mathbf{1}, \nu^{1,k}\rangle + \langle \mathbf{1}_{[0,\overline{\chi}_k(0)]}, \eta^{1,k}\rangle = \lambda_k(1-G_{1,k}^r(\chi_k))m_{1,k}^s + \lambda_k\int_0^{\chi_k}(1-G_{1,k}^r(x))dx$$

and

$$x_k^2 = \langle \mathbf{1}, \nu^{2,k}\rangle + \langle \mathbf{1}_{[0,\overline{\chi}_k(0)]}, \eta^{2,k}\rangle = w_k(1-G_{2,k}^r(\chi_k))m_{2,k}^s + w_k\int_0^{\chi_k}(1-G_{2,k}^r(x))dx.$$

Then $(x^1,x^2) \in \mathcal{X}_{(w,\chi)}$. This establishes (ii). Thus, the invariant state $(x^1,x^2,\nu^1,\nu^2,\eta^1,\eta^2) \in \mathcal{I}_\lambda$.

Claim 2. The set $\mathcal{I}_\lambda$ is a subset of invariant states. Fix $(w,\chi) \in \mathcal{Z}$ and $x = (x^1,x^2) \in \mathcal{X}_{(w,\chi)}$. Define $(\overline{X}^1,\overline{X}^2,\overline{\nu}^1,\overline{\nu}^2,\overline{\eta}^1,\overline{\eta}^2)$ to be such that for each $t \geq 0$,

$$(\overline{X}^1(t),\overline{X}^2(t),\overline{\nu}_t^1,\overline{\nu}_t^2,\overline{\eta}_t^1,\overline{\eta}_t^2) = (x^1,x^2,G^1(\chi)\mathrm{diag}(\lambda)\nu_*^1, G^2(\chi)\mathrm{diag}(w)\nu_*^2, \mathrm{diag}(\lambda)\eta_*^1, \mathrm{diag}(w)\eta_*^2).$$

We now show that $(\overline{X}^1,\overline{X}^2,\overline{\nu}^1,\overline{\nu}^2,\overline{\eta}^1,\overline{\eta}^2)$ is a solution to the fluid model equations associated with the initial data $(e_\lambda, x^1, x^2, G^1(\chi)\mathrm{diag}(\lambda)\nu_*^1, G^2(\chi)\mathrm{diag}(w)\nu_*^2, \mathrm{diag}(\lambda)\eta_*^1, \mathrm{diag}(w)\eta_*^2)$.

Fix $k \in \mathcal{K}$. Note that for each $t \geq 0$, $\overline{\nu}_t^{1,k} = \lambda_k(1 - G_{1,k}^r(\chi_k))\nu_*^{1,k}(dx)$ and $\overline{\nu}_t^{2,k} = w_k(1 - G_{2,k}^r(\chi_k))\nu_*^{2,k}(dx)$. It follows that for each $t \geq 0$,

$$\int_0^t \langle h_{1,k}^s, \overline{\nu}_u^{1,k} \rangle \, du = \lambda_k(1-G_{1,k}^r(\chi_k))t < \infty \text{ and } \int_0^t \langle h_{2,k}^s, \overline{\nu}_u^{2,k} \rangle \, du = w_k(1-G_{2,k}^r(\chi_k))t < \infty. \quad (4.28)$$

Similarly, note that for each $t \geq 0$, $\overline{\eta}_t^{1,k} = \lambda_k \eta_*^{1,k}(dx)$ and $\overline{\eta}_t^{2,k} = w_k \eta_*^{2,k}(dx)$, and then

$$\int_0^t \langle h_{1,k}^r, \overline{\eta}_u^{1,k} \rangle \, du = \lambda_k t < \infty \text{ and } \int_0^t \langle h_{2,k}^r, \overline{\eta}_u^{2,k} \rangle \, du = w_k t < \infty.$$

This establishes (2.5). For each $t \geq 0$, let

$$\begin{aligned} \bar{I}_k(t) &\doteq \sum_{l \in \mathcal{K}} P_{lk} \int_0^t \left(\langle h_{1,l}^s, \overline{\nu}_u^{1,l} \rangle + \langle h_{2,l}^s, \overline{\nu}_u^{2,l} \rangle \right) du & (4.29) \\ &= \sum_{l \in \mathcal{K}} P_{lk} \left(\lambda_k(1 - G_{1,k}^r(\chi_k)) + w_k(1 - G_{2,k}^r(\chi_k)) \right) t. & (4.30) \end{aligned}$$

Since $(w, \chi) \in \mathcal{Z}$, then (w, χ) satisfies (4.4), that is,

$$w = P' G^1(\chi)\lambda + P' G^2(\chi) w. \quad (4.31)$$

From the definitions of $G^1(\chi)$ and $G^2(\chi)$, it readily follows that

$$\sum_{l \in \mathcal{K}} P_{lk} \left(\lambda_k(1 - G_{1,k}^r(\chi_k)) + w_k(1 - G_{2,k}^r(\chi_k)) \right) = w_k.$$

Then, $\bar{I}_k(t) = w_k t$ for each $t \geq 0$. We now verify (2.10) and (2.11). For each $t \geq 0$ and each $f \in \mathcal{C}_b(\mathbb{R}_+)$,

$$\begin{aligned} & \int_{[0, H_{1,k}^r)} f(x+t) \frac{1 - G_{1,k}^r(x+t)}{1 - G_{1,k}^r(x)} \overline{\eta}_0^{1,k}(dx) + \int_{[0,t]} f(t-s)(1 - G_{1,k}^r(t-s))\lambda_k \, ds \\ = & \int_{[0, H_{1,k}^r)} f(x+t) \frac{1 - G_{1,k}^r(x+t)}{1 - G_{1,k}^r(x)} \lambda_k \, \eta_*^{1,k}(dx) + \int_{[0,t]} f(t-s)(1 - G_{1,k}^r(t-s))\lambda_k \, ds \\ = & \int_{[0, H_{1,k}^r)} f(x+t)(1 - G_{1,k}^r(x+t))\lambda_k \, dx + \int_{[0,t]} f(s)(1 - G_{1,k}^r(s))\lambda_k \, ds \\ = & \int_{[0, H_{1,k}^r)} f(x)(1 - G_{1,k}^r(x))\lambda_k \, dx = \int_{[0, H_{1,k}^r)} f(x) \, \eta_*^{1,k}(dx) = \int_{[0, H_{1,k}^r)} f(x) \, \overline{\eta}_t^{1,k}(dx). \end{aligned}$$

This establishes (2.10) and then (2.11) can be verified in a similar way using the fact that $\bar{I}_k(t) = w_k t$ for each $t \geq 0$. Note that for each $t \geq 0$,

$$\langle \mathbf{1}, \overline{\nu}_t^{1,k} \rangle = \lambda_k(1 - G_{1,k}^r(\chi_k)) \langle \mathbf{1}, \nu_*^{1,k} \rangle = \lambda_k(1 - G_{1,k}^r(\chi_k)) m_{1,k}^s,$$

and

$$\langle \mathbf{1}, \overline{\nu}_t^{2,k} \rangle = w_k(1 - G_{2,k}^r(\chi_k)) \langle \mathbf{1}, \nu_*^{2,k} \rangle = w_k(1 - G_{2,k}^r(\chi_k)) m_{2,k}^s.$$

Since $x = (x^1, x^2) \in \mathcal{X}_{(w,\chi)}$, we also have that

$$x_k^1 = \lambda_k(1 - G_{1,k}^r(\chi_k)) m_{1,k}^s + \lambda_k \int_0^{\chi_k} (1 - G_{1,k}^r(x)) dx,$$

and

$$x_k^2 = w_k(1 - G_{2,k}^r(\chi_k)) m_{2,k}^s + w_k \int_0^{\chi_k} (1 - G_{2,k}^r(x)) dx.$$

For each $t \geq 0$, define $\overline{\chi}_k(t) \doteq \chi_k$,

$$\overline{Q}^1_k(t) \doteq \overline{X}^1_k(t) - \langle \mathbf{1}, \overline{\nu}^{1,k}_t \rangle = x^1_k - \lambda_k(1 - G^r_{1,k}(\chi_k))m^s_{1,k} = \lambda_k \int_0^{\chi_k} (1 - G^r_{1,k}(x))dx,$$

and

$$\overline{Q}^2_k(t) \doteq \overline{X}^2_k(t) - \langle \mathbf{1}, \overline{\nu}^{2,k}_t \rangle = x^2_k - w_k(1 - G^r_{2,k}(\chi_k))m^s_{2,k} = w_k \int_0^{\chi_k} (1 - G^r_{2,k}(x))dx.$$

Note that for each $t \geq 0$,

$$\overline{\eta}^k_t \doteq \overline{\eta}^{1,k}_t + \overline{\eta}^{2,k}_t = \lambda_k \eta^{1,k}_* + w_k \eta^{2,k}_*.$$

Then it is readily to see that $\overline{\chi}_k(t) = \chi_k = (F^{\overline{\eta}^k_t})^{-1}\left(\overline{Q}^1_k(t) + \overline{Q}^2_k(t)\right)$ and then

$$\overline{Q}^1_k(t) = \langle \mathbf{1}_{[0,\overline{\chi}_k(t)]}, \overline{\eta}^{1,k}_t \rangle \text{ and } \overline{Q}^2_k(t) = \langle \mathbf{1}_{[0,\overline{\chi}_k(t)]}, \overline{\eta}^{2,k}_t \rangle.$$

This establishes (2.13) and (2.14). For each $t \geq 0$, define

$$\begin{aligned}
\overline{R}^1_k(t) &\doteq \int_0^t \left(\int_{[0,H^r_{1,k})} \mathbf{1}_{[0,\overline{\chi}_k(s)]}(u) h^r_{1,k}(u) \overline{\eta}^{1,k}_s(du) \right) ds = \lambda_k G^r_{1,k}(\chi_k)t, \\
\overline{R}^2_k(t) &\doteq \int_0^t \left(\int_{[0,H^r_{2,k})} \mathbf{1}_{[0,\overline{\chi}_k(s)]}(u) h^r_{2,k}(u) \overline{\eta}^{2,k}_s(du) \right) ds = w_k G^r_{2,k}(\chi_k)t.
\end{aligned}$$

This together with (4.28) implies that (2.17) and (2.18) hold. For each $t \geq 0$, define

$$\bar{L}^1_k(t) \doteq \lambda_k(1 - G^r_{1,k}(\chi_k))t \text{ and } \bar{L}^2_k(t) \doteq w_k(1 - G^r_{2,k}(\chi_k))t. \tag{4.32}$$

This implies that (2.6), (2.7), (2.8) and (2.9) hold. It remains to show that (2.19) holds. Note that (w, χ) also satisfies (4.5), that is,

$$\text{diag}(m^s_1)G^1(\chi)\lambda + \text{diag}(m^s_2)G^2(\chi)w \leq s.$$

In particular, $\lambda_k(1 - G^r_{1,k}(\chi_k))m^s_{1,k} + w_k(1 - G^r_{2,k}(\chi_k))m^s_{2,k} \leq s_k$. Note that for each $t \geq 0$,

$$\langle \mathbf{1}, \overline{\nu}^{1,k}_t \rangle + \langle \mathbf{1}, \overline{\nu}^{2,k}_t \rangle = \lambda_k(1 - G^r_{1,k}(\chi_k))m^s_{1,k} + w_k(1 - G^r_{2,k}(\chi_k))m^s_{2,k}.$$

Since $(w, \chi) \in \mathcal{Z}$, by the definition of $\mathcal{Z}$ in (4.6), there exists a set $\mathcal{J} \subseteq \mathcal{K}^\dagger$. We now consider two mutually exclusive cases:

1) $\lambda_k(1 - G^r_{1,k}(\chi_k))m^s_{1,k} + w_k(1 - G^r_{2,k}(\chi_k))m^s_{2,k} = s_k$. In this case, $s_k - \langle \mathbf{1}, \overline{\nu}^{1,k}_t \rangle - \langle \mathbf{1}, \overline{\nu}^{2,k}_t \rangle = 0$. On the other hand,

$$\overline{X}^1_k(t) + \overline{X}^2_k(t) = x^1_k + x^2_k = s_k + \lambda_k \int_0^{\chi_k} (1 - G^r_{1,k}(x))dx + w_k \int_0^{\chi_k} (1 - G^r_{2,k}(x))dx \geq s_k.$$

Then (2.19) holds in this case.

2) $\lambda_k(1 - G^r_{1,k}(\chi_k))m^s_{1,k} + w_k(1 - G^r_{2,k}(\chi_k))m^s_{2,k} < s_k$. In this case, we have that $k \notin \mathcal{J}$, and then $\lambda_k m^s_{1,k} + w_k m^s_{2,k} < s_k$ and $\chi_k = 0$. It follows that $x^1_k + x^2_k = \langle \mathbf{1}, \overline{\nu}^{1,k}_t \rangle + \langle \mathbf{1}, \overline{\nu}^{2,k}_t \rangle$ and then (2.19) also holds in this case.

Hence we have that $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ is a solution to the fluid model equations associated with the initial data $(e_\lambda, x^1, x^2, G^1(\chi)\text{diag}(\lambda)\nu^1_*, G^2(\chi)\text{diag}(w)\nu^2_*, \text{diag}(\lambda)\eta^1_*, \text{diag}(w)\eta^2_*)$. Thus $(x^1, x^2, G^1(\chi)\text{diag}(\lambda)\nu^1_*, G^2(\chi)\text{diag}(w)\nu^2_*, \text{diag}(\lambda)\eta^1_*, \text{diag}(w)\eta^2_*)$ is an invariant state. This established Claim 2 and completes the proof of Theorem 4.4. □

5. Convergence to the fluid model

In this section, we focus on the convergence of the stochastic evolution dynamics to the fluid model which we have studied. We first give some addition notation to be used. In §5.1, we give the representation of the system dynamics, and in §5.2, we prove the convergence of the fluid-scaled processes to the fluid model in the main paper.

Additional Notation. Given a Polish space $\mathcal{H}$, we denote by $\mathcal{D}_{\mathcal{H}}[0,T]$ (respectively, $\mathcal{D}_{\mathcal{H}}[0,\infty)$) the space of $\mathcal{H}$-valued, càdlàg functions on $[0,T]$ (respectively, $[0,\infty)$), and we endow this space with the usual Skorokhod J_1-topology [39],[47] so that they are Polish. A sequence $\{X_n\}$ of càdlàg, $\mathcal{H}$-valued processes, with X_n defined on $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$, is said to converge in distribution to a càdlàg $\mathcal{H}$-valued process X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ if, for every bounded, continuous functional $F : \mathcal{D}_{\mathcal{H}}[0,\infty) \to \mathbb{R}$, we have $\lim_{n\to\infty} \mathbb{E}_n [F(X_n)] = \mathbb{E}[F(X)]$, where $\mathbb{E}_n$ and $\mathbb{E}$ are the expectation operators with respect to the probability measures $\mathbb{P}_n$ and $\mathbb{P}$, respectively. Convergence in distribution of X_n to X will be denoted by $X_n \Rightarrow X$.

5.1. System Dynamics.

For each $k \in \mathcal{K}$, we shall use two measure-valued processes to describe the queue dynamics and another two measure-valued processes to describe the service dynamics, for externally arrived and internally routed customers of class k, respectively.

We first describe the potential queue dynamics for externally arrived customers of class k. For each $j \in \mathbb{Z}$, let $\zeta_j^{(N),1,k}$ denote the arrival time of external customer j of class k into the system. For $t \in [0,\infty)$, let $\eta_t^{(N),1,k}$ be a non-negative Borel measure on $[0, H^r_{1,k})$ with the representation:

$$\eta_t^{(N),1,k} = \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \delta_{w_j^{(N),1,k}(t)} \mathbb{1}_{\{w_j^{(N),1,k}(t) < r_j^{1,k}\}}, \tag{5.1}$$

where $w_j^{(N),1,k}(t) = \left[t - \zeta_j^{(N),1,k}\right] \vee 0 \wedge r_j^{1,k}$ represents the amount of time external customer j of class k has been in the potential queue by time t.

In a similar fashion, we can define another measure-valued process $\eta_t^{(N),2,k}$ to describe the potential queue dynamics for internally routed customers of class k. Specifically, let $\mathcal{C}_k^{(N)}$ be an a.s. $\mathbb{Z}_+$-valued random variable that represents the number of internal customers of class k that reentered the system due to internal routing by time zero and $\zeta_j^{(N),2,k}$ denote the time at which internal customer j of class k reenters the system upon service completion. Moreover, let $I_k^{(N)}(t)$ denote the cumulative number of internal customers of class k routed to the service station k in the time interval $(0,t]$. Since the service time distributions G_k^s, $k \in \mathcal{K}$, have densities on their supports, then with probability one, there is at most one customer finishes service from those K service stations at any given time $t \in [0,\infty)$, that is, $I_k^{(N)}(t) - I_k^{(N)}(t-) \in \{0,1\}$ for each $t \geq 0$ with probability one. Then $\eta_t^{(N),2,k}$ on $[0, H^r_{2,k})$ can be defined as

$$\eta_t^{(N),2,k} = \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \delta_{w_j^{(N),2,k}(t)} \mathbb{1}_{\{w_j^{(N),2,k}(t) < r_j^{2,k}\}}, \tag{5.2}$$

where $w_j^{(N),2,k}(t) = \left[t - \zeta_j^{(N),2,k}\right] \vee 0 \wedge r_j^{2,k}$ represents the amount of time internal customer j of class k has been in the potential queue by time t.

For each $t \geq 0$, let

$$\eta_t^{(N),k} = \eta_t^{(N),1,k} + \eta_t^{(N),2,k}. \tag{5.3}$$

The measure $\eta_t^{(N),k}$ keeps track of the waiting times of all customers in the potential queue at time t. Note that $\langle \mathbf{1}, \eta_t^{(N),1,k} \rangle = \eta_t^{(N),1,k}[0,\infty)$ represents the total number of external customers waiting

in the potential queue at time t, and $\langle \mathbf{1}, \eta_t^{(N),2,k} \rangle = \eta_t^{(N),2,k}[0,\infty)$ represents the total number of internal customers waiting in the potential queue at time t. Thus, $\langle \mathbf{1}, \eta_t^{(N),k} \rangle = \eta_t^{(N),k}[0,\infty)$ represents the total number of customers waiting in the potential queue at time t.

For $t \in [0,\infty)$, let $X_k^{(N),1}(t)$ be the total number of external customers of class k in the system and $Q_k^{(N),1}(t)$ be the number of external customers of class k waiting in queue at time t. Similarly, let $X_k^{(N),2}(t)$ be the total number of internal customers of class k in the system and $Q_k^{(N),2}(t)$ be the number of internal customers of class k waiting in queue at time t. Due to the non-idling condition, the queue length process $Q_k^{(N),1} + Q_k^{(N),2}$ of external and internal customers of class k is then given by

$$Q_k^{(N),1}(t) + Q^{(N),2}(t) = [X_k^{(N),1}(t) + X_k^{(N),2}(t) - N_k]^+. \tag{5.4}$$

Moreover, since the head-of-the-line customer (external or internal) of class k in queue is the customer of class k in queue with the longest waiting time, the quantity

$$\chi_k^{(N)}(t) \doteq \inf\left\{x > 0 : \ \eta_t^{(N),k}[0,x] \geq Q_k^{(N),1}(t) + Q_k^{(N),2}(t)\right\} \tag{5.5}$$

represents the waiting time of the head-of-the-line customer of class k in the queue at time t. Since this is an FCFS system, any mass in $\eta_t^{(N),k}$ that lies to the right of $\chi_k^{(N)}(t)$ represents a customer that is either in service or has departed by time t. Therefore, the queue length process $Q_k^{(N),1}$ of external customers and $Q_k^{(N),2}$ of internal customers admit the following alternative representation in terms of $\chi_k^{(N)}$, $\eta^{(N),1,k}$ and $\eta^{(N),2,k}$:

$$Q_k^{(N),1}(t) = \eta_t^{(N),1,k}[0, \chi_k^{(N)}(t)] \text{ and } Q_k^{(N),2}(t) = \eta_t^{(N),2,k}[0, \chi_k^{(N)}(t)]. \tag{5.6}$$

For $t \in [0,\infty)$, in a fashion analogous to (5.1) and (5.2), we can also define $\nu_t^{(N),1,k}$ to be a discrete non-negative Borel measure on $[0, H_{1,k}^s)$ that has a unit mass at the amount of time each of the externally arrived customers has spent in service by time t, and similarly, $\nu_t^{(N),2,k}$ a discrete non-negative Borel measure on $[0, H_{2,k}^s)$ that has a unit mass at the amount of time each of the internally routed customers has spent in service by time t. For each $j \in \mathbb{Z}$, let $\varsigma_j^{(N),1,k}$ denote the time at which external customer j of class k enters service and $\varsigma_j^{(N),2,k}$ denote the time at which internal customer j of class k enters service. Note that if external (resp. internal) customer j of class k reneged, then $\varsigma_j^{(N),1,k} = \infty$ (resp. $\varsigma_j^{(N),2,k} = \infty$). Let $a_j^{(N),1,k}(t) = \left[t - \varsigma_j^{(N),1,k}\right] \vee 0 \wedge v_j^{1,k}$ (resp. $a_j^{(N),2,k}(t) = \left[t - \varsigma_j^{(N),2,k}\right] \vee 0 \wedge v_j^{2,k}$) represents the amount of time external (resp. internal) customer j of class k has been in service by time t. Then, we can write

$$\nu_t^{(N),1,k} = \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \delta_{a_j^{(N),1,k}(t)} \mathbb{1}_{\left\{a_j^{(N),1,k} < v_j^{1,k}\right\}} \text{ and } \nu_t^{(N),2,k} = \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \delta_{a_j^{(N),2,k}(t)} \mathbb{1}_{\left\{a_j^{(N),2,k} < v_j^{2,k}\right\}}.$$

Also, define $\nu_t^{(N),k}$ by

$$\nu_t^{(N),k} \quad = \quad \nu_t^{(N),1,k} + \nu_t^{(N),2,k}, \tag{5.7}$$

Note that $\langle \mathbf{1}, \nu_t^{(N),k} \rangle = \nu_t^{(N),k}[0,\infty)$ represents the total number of customers of class k in service at time t.

We now introduce some auxiliary processes. Fix $k \in \mathcal{K}$. Let $D_{kl}^{(N),1}$ (resp. $D_{kl}^{(N),2}$), $l \in \mathcal{K} \cup \{0\}$, denote the cumulative routing processes of external (resp. internal) customers, where for each $l \in \mathcal{K}$, $D_{kl}^{(N),1}(t)$ (resp. $D_{kl}^{(N),2}(t)$) is the cumulative number of external (resp. internal) customers of class k that have completed the service and joined class l in the time interval $[0,t]$, and $D_{k0}^{(N),1}(t)$

(resp. $D_{k0}^{(N),2}(t)$) is the cumulative number of external (resp. internal) customers of class k that have completed the service and left the system in the interval $[0,t]$. Then $D_{kl}^{(N),1}(t)$ and $D_{kl}^{(N),2}(t)$ have the representations

$$D_{kl}^{(N),1}(t) = \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\{\phi^{1,k}(j)=e_l\}} \mathbb{1}_{\left\{\frac{da_j^{(N),1,k}}{dt}(s-)>0,\ \frac{da_j^{(N),1,k}}{dt}(s+)=0\right\}}, \tag{5.8}$$

$$D_{kl}^{(N),2}(t) = \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\{\phi^{2,k}(j)=e_l\}} \mathbb{1}_{\left\{\frac{da_j^{(N),2,k}}{dt}(s-)>0,\ \frac{da_j^{(N),2,k}}{dt}(s+)=0\right\}}. \tag{5.9}$$

It is obvious that

$$I_k^{(N)}(t) = \sum_{l\in\mathcal{K}} (D_{lk}^{(N),1}(t) + D_{lk}^{(N),2}(t)). \tag{5.10}$$

In addition, the departure process $D_k^{(N),1}$ (resp. $D_k^{(N),2}$), where $D_k^{(N),1}(t)$ (resp. $D_k^{(N),2}(t)$) represents the cumulative number of external (resp. internal) customers of class k that have completed service from the service station k in the time interval $[0,t]$, can be represented in term of $D_{kl}^{(N),1}$ (resp. $D_{kl}^{(N),2}$), $l \in \mathcal{K}\cup\{0\}$, as

$$D_k^{(N),1}(t) = \sum_{l\in\mathcal{K}\cup\{0\}} D_{kl}^{(N),1}(t) \text{ and } D_k^{(N),2}(t) = \sum_{l\in\mathcal{K}\cup\{0\}} D_{kl}^{(N),2}(t). \tag{5.11}$$

Let $S_k^{(N),1}$ (resp. $S_k^{(N),2}$) denote the cumulative potential reneging process, where $S_k^{(N),1}(t)$ (resp. $S_k^{(N),2}(t)$) represents the cumulative number of external (resp. internal) customers of class k whose waiting times in the potential queue have reached their patience times in the interval $[0,t]$. Thus, $S_k^{(N),1}$ (resp. $S_k^{(N),2}$) admits the representation

$$S_k^{(N),1}(t) = \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw_j^{(N),1,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),1,k}}{dt}(s+)=0\right\}}, \tag{5.12}$$

and

$$S_k^{(N),2}(t) = \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw_j^{(N),2,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),2,k}}{dt}(s+)=0\right\}}. \tag{5.13}$$

Let $R_k^{(N),1}$ (resp. $R_k^{(N),2}$) denote the cumulative reneging process, where $R_k^{(N),1}(t)$ (resp. $R_k^{(N),2}(t)$) is the cumulative number of external (resp. inetrnal) customers of class k that have reneged in the time interval $[0,t]$. Then $R_k^{(N),1}$ (resp. $R_k^{(N),2}$) admit the representation

$$R_k^{(N),1}(t) = \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{w_j^{(N),1,k}(s)\le\chi_k^{(N)}(s-),\frac{dw_j^{(N),1,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),1,k}}{dt}(s+)=0\right\}}, \tag{5.14}$$

and

$$R_k^{(N),2}(t) = \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{w_j^{(N),2,k}(s)\le\chi_k^{(N)}(s-),\frac{dw_j^{(N),2,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),2,k}}{dt}(s+)=0\right\}}, \tag{5.15}$$

where the additional restrictions $w_j^{(N),1,k}(s) \le \chi_k^{(N)}(s-)$ and $w_j^{(N),2,k}(s) \le \chi_k^{(N)}(s-)$ are imposed so as to only count the reneging of customers of class k (including external and internal customers) actually in queue. Here, one considers the left limit $\chi_k^{(N)}(s-)$ of $\chi_k^{(N)}$ at time s to capture the situation in which $\chi_k^{(N)}$ jumps down at time s due to the head-of-the-line customer of class k reneging from the queue or entering service.

Therefore, for each $k \in \mathcal{K}$, the mass balances on the total numbers of external customers and internal customers of class k in the system, the numbers of external customers and internal customers of class k waiting in the "potential queue", and the numbers of external customers and internal customers of class k in service, show that

$$X_k^{(N),1}(0) + E_k^{(N)} = X_k^{(N),1} + R_k^{(N),1} + D_k^{(N),1}, \tag{5.16}$$

$$X_k^{(N),2}(0) + I_k^{(N)} = X_k^{(N),2} + R_k^{(N),2} + D_k^{(N),2}, \tag{5.17}$$

$$\langle \mathbf{1}, \eta_0^{(N),1,k} \rangle + E_k^{(N)} = \langle \mathbf{1}, \eta^{(N),1,k} \rangle + S_k^{(N),1}, \tag{5.18}$$

$$\langle \mathbf{1}, \eta_0^{(N),2,k} \rangle + I_k^{(N)} = \langle \mathbf{1}, \eta^{(N),2,k} \rangle + S_k^{(N),2}, \tag{5.19}$$

$$\langle \mathbf{1}, \nu_0^{(N),1,k} \rangle + L_k^{(N),1} = \langle \mathbf{1}, \nu^{(N),1,k} \rangle + D_k^{(N),1}, \tag{5.20}$$

$$\langle \mathbf{1}, \nu_0^{(N),2,k} \rangle + L_k^{(N),2} = \langle \mathbf{1}, \nu^{(N),2,k} \rangle + D_k^{(N),2}, \tag{5.21}$$

where $L_k^{(N),1}(t)$ (resp. $L_k^{(N),2}(t)$) represents the cumulative number of external (resp. internal) customers of class k that have entered service in the interval $[0,t]$. In addition, it is also clear that

$$X_k^{(N),1} = \langle \mathbf{1}, \nu^{(N),1,k} \rangle + Q_k^{(N),1}. \tag{5.22}$$

$$X_k^{(N),2} = \langle \mathbf{1}, \nu^{(N),2,k} \rangle + Q_k^{(N),2}. \tag{5.23}$$

Combining (5.16), (5.20), (5.22), (5.17), (5.21) and (5.23), we obtain the following mass balance equation for the numbers of external and internal customers in queue, respectively:

$$Q_k^{(N),1}(0) + E_k^{(N)} = Q_k^{(N),1} + R_k^{(N),1} + L_k^{(N),1}, \tag{5.24}$$

$$Q_k^{(N),2}(0) + I_k^{(N)} = Q_k^{(N),2} + R_k^{(N),2} + L_k^{(N),2}. \tag{5.25}$$

Furthermore, the non-idling condition takes the form

$$N_k - \langle \mathbf{1}, \nu^{(N),1,k} \rangle - \langle \mathbf{1}, \nu^{(N),2,k} \rangle = [N_k - X_k^{(N),1} - X_k^{(N),2}]^+. \tag{5.26}$$

Note that if $N_k = \infty$, then the above non-idling condition holds automatically.

5.2. **Proof for the convergence.** Consider the following scaled versions of the processes described above. For each $N \in \mathbb{N}$, the scaled version of the state descriptor $(\overline{E}^{(N)}, \overline{X}^{(N),1}, \overline{X}^{(N),2}, \overline{\nu}^{(N),1}, \overline{\nu}^{(N),2}, \overline{\eta}^{(N),1}, \overline{\eta}^{(N),2})$ is given by

$$\begin{aligned} \overline{E}^{(N)}(t) &\doteq \frac{E^{(N)}(t)}{N}, \quad \overline{X}^{(N),1}(t) \doteq \frac{X^{(N),1}(t)}{N}, \quad \overline{X}^{(N),2}(t) \doteq \frac{X^{(N),2}(t)}{N}, \overline{\nu}_t^{(N),1}(B) \doteq \frac{\nu_t^{(N),1}(B)}{N}, \\ \overline{\nu}_t^{(N),2}(B) &\doteq \frac{\nu_t^{(N),2}(B)}{N}, \overline{\eta}_t^{(N),1}(B) \doteq \frac{\eta_t^{(N),1}(B)}{N}, \qquad \overline{\eta}_t^{(N),2}(B) \doteq \frac{\eta_t^{(N),2}(B)}{N} \end{aligned} \tag{5.27}$$

for $t \in [0,\infty)$ and any Borel subset B of $\mathbb{R}_+$. Analogously, define

$$\overline{A}^{(N)} \doteq \frac{A^{(N)}}{N} \text{ for } A = D, L, Q, R, S, I. \tag{5.28}$$

Our goal is to identify the limit in distribution of the quantities $(\overline{X}^{(N),1}, \overline{X}^{(N),2}, \overline{\nu}^{(N),1}, \overline{\nu}^{(N),2}, \overline{\eta}^{(N),1}, \overline{\eta}^{(N),2})$, as $N \to \infty$. To this end, we impose some natural assumptions on the sequence of initial conditions $(\overline{E}^{(N)}, \overline{X}^{(N),1}(0), \overline{X}^{(N),2}(0), \overline{\nu}_0^{(N),1}, \overline{\nu}_0^{(N),2}, \overline{\eta}_0^{(N),1}, \overline{\eta}_0^{(N),2})$.

Assumption 5.1. (Initial conditions) *There exists an $\mathcal{S}_0$-valued random variable $(\overline{E}, \overline{X}(0), \overline{\nu}_0^1, \overline{\nu}_0^2, \overline{\eta}_0^1, \overline{\eta}_0^2)$ such that, as $N \to \infty$, the following limits hold $\mathbb{P}$-a.s.:*

(i) $\overline{E}^{(N)} \to \overline{E}$ in $\mathcal{D}_{\mathbb{R}_+}[0,\infty)^K$, where $\overline{E}$ is continuous, $\overline{E}(0) = 0$, and $\mathbb{E}\left[\overline{E}^{(N)}(t)\right] \to \mathbb{E}\left[\overline{E}(t)\right] < \infty$ for every $t \in [0,\infty)$;

(ii) $\overline{X}^{(N),1}(0) \to \overline{X}^1(0)$ and $\overline{X}^{(N),2}(0) \to \overline{X}^2(0)$ in $\mathbb{R}_+^K$, and $\mathbb{E}[\overline{X}^{(N),1}(0)] \to \mathbb{E}[\overline{X}^1(0)]$ and $\mathbb{E}[\overline{X}^{(N),2}(0)] \to \mathbb{E}[\overline{X}^2(0)]$ in $\mathbb{R}_+^K$;

(iii) $\overline{\nu}_0^{(N),j} \xrightarrow{w} \overline{\nu}_0^j$ in $\Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H_{j,k}^s)$, $j = 1, 2$;

(iv) $\overline{\eta}_0^{(N),j} \xrightarrow{w} \overline{\eta}_0^j$ in $\Pi_{k\in\mathcal{K}}\mathcal{M}_F[0, H_{j,k}^r)$, where $\overline{\eta}_0^j$ is continuous on $\mathbb{R}_+$, and $\mathbb{E}\left[\langle \mathbf{1}, \overline{\eta}_0^{(N),j}\rangle\right] \to \mathbb{E}[\langle \mathbf{1}, \overline{\eta}_0^j\rangle] < \infty$, for $j = 1, 2$.

In order to establish the convergence result, we impose the following assumptions on $G_{j,k}^r$ and G_k^s, $j = 1, 2$ and $k \in \mathcal{K}$.

Assumption 5.2. *For each $k \in \mathcal{K}$ and $j = 1, 2$, there exists $L_{j,k}^s < H_{j,k}^s$ such that $h_{j,k}^s$ is either bounded or lower-semicontinuous on $(L_{j,k}^s, H_{j,k}^s)$, $h_{j,k}^r$ is locally bounded and there exists $L_{j,k}^r < H_{j,k}^r$ such that $h_{j,k}^r$ is either bounded or lower-semicontinuous on $(L_{j,k}^r, H_{j,k}^r)$.*

Theorem 5.1. *Suppose that Assumptions 3.1–3.3 and 5.1–5.2 hold. Let $(\overline{E}, \overline{X}^1(0), \overline{X}^2(0)\overline{\nu}_0^1, \overline{\nu}_0^2, \overline{\eta}_0^1, \overline{\eta}_0^2) \in \mathcal{S}_0$ be the limiting initial condition. Then there exists a unique solution $(\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$ to the associated fluid equations (2.5) – (2.19), and*

$$(\overline{X}^{(N),1}, \overline{X}^{(N),2}, \overline{\nu}^{(N),1}, \overline{\nu}^{(N),2}, \overline{\eta}^{(N),1}, \overline{\eta}^{(N),2}) \Rightarrow (\overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2) \quad as \quad N \to \infty. \tag{5.29}$$

The proof of this theorem can be carried out in two steps as the proof of Theorem 3.6 of [16]. The first step, stated in Theorem 5.2, is to show the fluid-scaled processes are tight and the second step, stated in Theorem 5.5, is to show that every limit of any subsequence of the fluid-scaled processes solves the fluid equations.

To state Theorem 5.2, we first introduce some additional processes that are used in proving the convergence. For each $k \in \mathcal{K}$ and any measurable function φ on $[0, H_{1,k}^s) \times \mathbb{R}_+$, consider the process $D_\varphi^{(N),1,k}$ that takes values in $\mathbb{R}$, and is given by

$$D_\varphi^{(N),1,k}(t) \doteq \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{da_j^{(N),1,k}}{dt}(s-)>0,\ \frac{da_j^{(N),1,k}}{dt}(s+)=0\right\}} \varphi(a_j^{(N),1,k}(s), s), \tag{5.30}$$

for $t \in [0,\infty)$ and for any measurable function φ on $[0, H_{2,k}^s) \times \mathbb{R}_+$, consider the process $D_\varphi^{(N),2,k}$ that takes values in $\mathbb{R}$, and is given by

$$D_\varphi^{(N),2,k}(t) \doteq \sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{da_j^{(N),2,k}}{dt}(s-)>0,\ \frac{da_j^{(N),2,k}}{dt}(s+)=0\right\}} \varphi(a_j^{(N),2,k}(s), s), \tag{5.31}$$

for $t \in [0,\infty)$. In an exactly analogous fashion, for each $k \in \mathcal{K}$, any measurable function φ on $[0, H_{1,k}^r) \times \mathbb{R}_+$, consider the process $S_\varphi^{(N),1,k}$ that takes values in $\mathbb{R}$, and is given by

$$S_\varphi^{(N),1,k}(t) \doteq \sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw_j^{(N),1,k}}{dt}(s-)>0,\ \frac{dw_j^{(N),1,k}}{dt}(s+)=0\right\}} \varphi(w_j^{(N),1,k}(s), s), \tag{5.32}$$

for $t \in [0,\infty)$, and for any measurable function φ on $[0, H^r_{2,k}) \times \mathbb{R}_+$, consider the process $S^{(N),2,k}_\varphi$ that takes values in $\mathbb{R}$, and is given by

$$S^{(N),2,k}_\varphi(t) \doteq \sum_{j=-\mathcal{C}^{(N)}_k+1}^{I^{(N)}_k(t)} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{dw^{(N),2,k}_j}{dt}(s-)>0,\ \frac{dw^{(N),2,k}_j}{dt}(s+)=0\right\}} \varphi(w^{(N),2,k}_j(s), s), \tag{5.33}$$

for $t \in [0,\infty)$. Next, comparing (5.14) with (5.32) and (5.15) with (5.33), it is clear that for each $k \in \mathcal{K}$,

$$R^{(N),1}_k(t) = S^{(N),1,k}_{\theta^{(N)}_k}(t) \text{ and } R^{(N),2}(t) = S^{(N),2,k}_{\theta^{(N)}_k}(t), \qquad t \geq 0, \tag{5.34}$$

where $\theta^{(N)}_k$ is given by

$$\theta^{(N)}_k(x,s) = \mathbb{1}_{[0,\chi^{(N)}_k(s-)]}(x), \qquad x \in \mathbb{R},\ s \geq 0. \tag{5.35}$$

For $t \in [0,\infty)$, let $\tilde{\mathcal{F}}^{(N)}_t$ be the σ-algebra generated by

$$\left\{ \begin{array}{c} \mathcal{E}^{(N)}_k, \mathcal{C}^{(N)}_k, X^{(N),1}_k(0), X^{(N),2}_k(0), \alpha^{(N)}_{E_k}(s), w^{(N),1,k}_i(s), w^{(N),2,k}_j(s), a^{(N),1,k}_i(s), a^{(N),2,k}_j(s), \\ s^{(N),1,k}_i(s), s^{(N),2,k}_j(s) : i \in \{-\mathcal{E}^{(N)}_k+1,\dots,0\} \cup \mathbb{N}, j \in \{-\mathcal{C}^{(N)}_k+1,\dots,0\} \cup \mathbb{N}, \\ \phi^{1,k}(l),\ -\mathcal{E}^{(N)}_k + 1 \leq l \leq \max\left|\{n :\ a^{(N),1,k}_n(s) > 0\}\right|, \\ \phi^{2,k}(l),\ -\mathcal{C}^{(N)}_k + 1 \leq l \leq \max\left|\{n :\ a^{(N),2,k}_n(s) > 0\}\right|,\ s \in [0,t], k \in \mathcal{K} \end{array} \right\},$$

where $s^{(N),1,k}_i(s)$ is equal to the index of the station at which the external customer i of class k receives/received service if it has already entered service by time s and $s^{(N),1,k}_i(s) = 0$ otherwise, and $s^{(N),2,k}_j(s)$ is equal to the index of the station at which the internal customer j of class k receives/received service if it has already entered service by time s and $s^{(N),2,k}_j(s) = 0$ otherwise, and let $\{\mathcal{F}^{(N)}_t\}$ denote the associated right-continuous filtration, completed with respect to $\mathbb{P}$. By using a similar construction as in Appendix A of [16], we can see that all the processes $E^{(N)}, X^{(N),1}, X^{(N),2}, \nu^{(N),1}, \nu^{(N),2}, \eta^{(N),1}, \eta^{(N),2}$ and the auxiliary processes are $\{\mathcal{F}^{(N)}_t\}$-adapted. It follows immediately from (5.30),(5.31), (5.32), (5.33) and the right continuity of the filtration $\{\mathcal{F}^{(N)}_t\}$ that for each $k \in \mathcal{K}$, $D^{(N),1,k}_\varphi$, $D^{(N),2,k}_\varphi$, $S^{(N),1,k}_\varphi$ and $S^{(N),2,k}_\varphi$ are $\{\mathcal{F}^{(N)}_t\}$-adapted.

Fix $k \in \mathcal{K}$. For for each $j = 1,2$ and any bounded measurable function φ on $[0, H^s_{j,k}) \times \mathbb{R}_+$, consider the sequence $\{A^{(N),j,k}_{\varphi,\nu}\}$ of processes given by

$$A^{(N),j,k}_{\varphi,\nu}(t) \doteq \int_0^t \left(\int_{[0,H^s_{j,k})} \varphi(x,s) h^s_{j,k}(x)\, \nu^{(N),j,k}_s(dx) \right) ds, \qquad t \in [0,\infty). \tag{5.36}$$

Likewise, for each $j = 1,2$ and any bounded measurable function φ on $[0, H^r_{j,k}) \times \mathbb{R}_+$, let

$$A^{(N),j,k}_{\varphi,\eta}(t) \doteq \int_0^t \left(\int_{[0,H^r_{j,k})} \varphi(x,s) h^r_{j,k}(x)\, \eta^{(N),j,k}_s(dx) \right) ds, \qquad t \in [0,\infty), \tag{5.37}$$

and

$$A^{(N),j,k}_{\theta^{(N)}_k,\eta}(t) \doteq \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\chi^{(N)}_k(s-)]}(x) h^r_{j,k}(x)\, \eta^{(N),j,k}_s(dx) \right) ds, \qquad t \in [0,\infty), \tag{5.38}$$

where $\theta^{(N)}_k$ is defined in (5.35). A similar argument as Proposition 5.1 and Lemma 5.4 of [16] shows that for each $j = 1,2$, $A^{(N),j,k}_{\varphi,\nu}$ (respectively, $A^{(N),j,k}_{\varphi,\eta}$) is the $\mathcal{F}^{(N)}_t$-compensator of process $D^{(N),j.k}_\varphi$

(respectively, $S_{\varphi}^{(N),j,k}$) and $A_{\theta_k^{(N)},\eta}^{(N),1,k} + A_{\theta_k^{(N)},\eta}^{(N),2,k}$ is the $\mathcal{F}_t^{(N)}$-compensator of process $R_k^{(N)}$. That is, for each $k \in \mathcal{K}$ and $j = 1, 2$, and for every bounded measurable function φ on $[0, H_{j,k}^s) \times \mathbb{R}_+$ such that the function $s \mapsto \varphi(a_i^{(N),j,k}(s), s)$ is left continuous on $[0, \infty)$ for each $j = 1, 2$ and $i \in \mathbb{Z}$, the process $M_{\varphi,\nu}^{(N),j,k}$ defined by

$$M_{\varphi,\nu}^{(N),j,k} \doteq D_{\varphi}^{(N),j,k} - A_{\varphi,\nu}^{(N),j,k} \tag{5.39}$$

is a local $\mathcal{F}_t^{(N)}$-martingale. Moreover, for each $j = 1, 2$, $N \in \mathbb{N}$, $t \in [0, \infty)$ and $m \in [0, H_{j,k}^s)$,

$$|A_{\varphi,\nu}^{(N),j,k}(t)| \le \|\varphi\|_{\infty} \left(X_k^{(N)}(0) + E_k^{(N)}(t) + I_k^{(N)}(t)\right) \left(\int_0^m h_{j,k}^s(x)\, dx\right) < \infty \tag{5.40}$$

for every $\varphi \in \mathcal{C}_c([0, H_{j,k}^s) \times \mathbb{R}_+)$ with $\text{supp}(\varphi) \subset [0, m] \times \mathbb{R}_+$. In addition, the quadratic variation process $\langle \overline{M}_{\varphi,\nu}^{(N),j,k} \rangle$ of the scaled process $\overline{M}_{\varphi,\nu}^{(N),j,k} \doteq M_{\varphi,\nu}^{(N),j,k}/N$ satisfies

$$\lim_{N\to\infty} \mathbb{E}\left[\langle \overline{M}_{\varphi,\nu}^{(N),j,k} \rangle(t)\right] = 0; \qquad \overline{M}_{\varphi,\nu}^{(N),j,k} \Rightarrow \mathbf{0} \text{ as } N \to \infty. \tag{5.41}$$

Furthermore, properties (5.39)–(5.41) also hold with $D^{(N),j,k}, A^{(N),j,k}, M^{(N),j,k}, a, \nu, H_{j,k}^s$ and $h_{j,k}^s$, respectively, replaced by $S^{(N),j,k}, A^{(N),j,k}, M^{(N),j,k}, w, \eta, H_{j,k}^r$ and $h_{j,k}^r$ for $j = 1, 2$. Also the processes $M_{\theta_k^{(N)},\eta}^{(N),1,k}$ and $M_{\theta_k^{(N)},\eta}^{(N),2,k}$ defined by

$$M_{\theta_k^{(N)},\eta}^{(N),1,k} \doteq R_k^{(N),1} - A_{\theta_k^{(N)},\eta}^{(N),1,k} \quad \text{and} \quad M_{\theta_k^{(N)},\eta}^{(N),2,k} \doteq R_k^{(N),2} - A_{\theta_k^{(N)},\eta}^{(N),2,k} \tag{5.42}$$

are local $\mathcal{F}_t^{(N)}$-martingales and satisfy for $j = 1, 2$, as $N \to \infty$,

$$\lim_{N\to\infty} \mathbb{E}\left[\langle \overline{M}_{\theta_k^{(N)},\eta}^{(N),j,k} \rangle(t)\right] = 0 \quad \text{and} \quad \overline{M}_{\theta_k^{(N)},\eta}^{(N),j,k} \Rightarrow \mathbf{0}. \tag{5.43}$$

Notice that by (5.8), (5.9) and (5.10), we have for each $k \in \mathcal{K}$, $l \in \mathcal{K} \cup \{0\}$ and $t \ge 0$,

$$\begin{aligned}
\mathbb{E}\left[D_{lk}^{(N),1}(t) + D_{lk}^{(N),2}(t)\right] &\le \mathbb{E}\left[\sum_{j=1}^{E_l^{(N)}(t)} \mathbb{1}_{\{\phi^{1,l}(j)=e_k\}}\right] + \mathbb{E}\left[\sum_{j=1}^{I_l^{(N)}(t)} \mathbb{1}_{\{\phi^{2,l}(j)=e_k\}}\right] \\
&\quad + \mathbb{E}\left[\sum_{j=-\mathcal{E}_l^{(N)}+1}^{0} \mathbb{1}_{\{\phi^{1,l}(j)=e_k\}} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{da_j^{(N),1,l}}{dt}(0+)>0\right\}}\right] \\
&\quad + \mathbb{E}\left[\sum_{j=-\mathcal{C}_l^{(N)}+1}^{0} \mathbb{1}_{\{\phi^{2,l}(j)=e_k\}} \sum_{s\in[0,t]} \mathbb{1}_{\left\{\frac{da_j^{(N),2,l}}{dt}(0+)>0\right\}}\right] \\
&\le P_{lk}\mathbb{E}\left[E_l^{(N)}(t) + I_l^{(N)}(t) + X_l^{(N),1}(0) + X_l^{(N),2}(0)\right],
\end{aligned}$$

and

$$\mathbb{E}\left[I_k^{(N)}(t)\right] = \sum_{l\in\mathcal{K}} \mathbb{E}\left[D_{lk}^{(N),1}(t) + D_{lk}^{(N),2}(t)\right] \le \sum_{l\in\mathcal{K}} P_{lk}\mathbb{E}\left[E_l^{(N)}(t) + I_l^{(N)}(t) + X_l^{(N),1}(0) + X_l^{(N),2}(0)\right].$$

Since the above inequality holds for each $k \in \mathcal{K}$, by treating it in the vector form and using the fact that $H = (I - P')^{-1}$ has non-negative entries, we have that

$$\mathbb{E}\left[I_k^{(N)}(t)\right] \le \sum_{l\in\mathcal{K}} (HP')_{kl}\mathbb{E}\left[E_l^{(N)}(t) + X_l^{(N),1}(0) + X_l^{(N),2}(0)\right] < \infty, \tag{5.44}$$

where the last inequality holds due to Assumption 5.1. In addition, using (5.16), (5.22), (5.17), (5.23), (5.44) and the non-negativity of $Q_k^{(N)}$, $R_k^{(N)}$ and $\langle \mathbf{1}, \nu^{(N),k} \rangle$, it follows from Assumption 5.1 that for $j = 1, 2$, any $t \in [0, \infty)$ and any bounded, measurable φ,

$$\mathbb{E}\left[\left|D_\varphi^{(N),j,k}(t)\right|\right] \leq \|\varphi\|_\infty \, \mathbb{E}\left[X_k^{(N),1}(0) + X_k^{(N),2}(0) + E_k^{(N)}(t) + I_k^{(N)}(t)\right] < \infty \tag{5.45}$$

and likewise, for each $t \in [0, \infty)$ and bounded measurable φ and ψ, (5.18) and (5.19) show that

$$\mathbb{E}\left[\left|S_\varphi^{(N),1,k}(t)\right| + \left|S_\psi^{(N),2,k}(t)\right|\right] \leq (\|\psi\|_\infty + \|\varphi\|_\infty)\mathbb{E}\left[\langle \mathbf{1}, \eta_0^{(N)} \rangle + E_k^{(N)}(t) + I_k^{(N)}(t)\right] < \infty. \tag{5.46}$$

From (5.45) and (5.40) it is clear that for each $j = 1, 2$, $t \in [0, \infty)$, the linear functionals $\overline{D}_\cdot^{(N),j,k}(t) : \varphi \mapsto \overline{D}_\varphi^{(N),j,k}(t)$ and $\overline{A}_{\cdot,\nu}^{(N),j,k}(t) : \varphi \mapsto \overline{A}_{\varphi,\nu}^{(N),j,k}(t)$ are finite Radon measures on $[0, H_{j,k}^s) \times \mathbb{R}_+$. Likewise, from (5.46) and the fact that (5.40) holds with $\nu^{(N),j,k}$, $h_{j,k}^s$, respectively, replaced by $\eta^{(N),j,k}$, $h_{j,k}^r$, $j = 1, 2$, it follows that for each $j = 1, 2$, the linear functionals $\overline{S}_\cdot^{(N),j,k}(t) : \varphi \mapsto \overline{S}_\varphi^{(N),j,k}(t)$ and $\overline{A}_{\cdot,\eta}^{(N),j,k}(t) : \varphi \mapsto \overline{A}_{\varphi,\eta}^{(N),j,k}(t)$ define finite Radon measures on $[0, H_{j,k}^r) \times \mathbb{R}_+$. Thus $\{\overline{D}_\cdot^{(N),j,k}(t) : t \in [0, \infty)\}$ and $\{\overline{A}_{\cdot,\nu}^{(N),j,k}(t) : t \in [0, \infty)\}$ can be viewed as $\mathcal{M}_F([0, H_{j,k}^s) \times \mathbb{R}_+)$-valued càdlàg processes, and $\{\overline{S}_\cdot^{(N),j,k}(t) : t \in [0, \infty)\}$ and $\{\overline{A}_{\cdot,\eta}^{(N),j,k}(t) : t \in [0, \infty)\}$ can be viewed as $\mathcal{M}_F([0, H_{j,k}^r) \times \mathbb{R}_+)$-valued càdlàg processes for $j = 1, 2$. Now, for each $N \in \mathbb{N}$ and $k \in \mathcal{K}$, let

$$\begin{aligned}
\overline{Z}_k^{(N)} \doteq \Big(&\overline{X}_k^{(N),1}(0), \overline{X}_k^{(N),2}(0), \overline{E}_k^{(N)}, \overline{X}_k^{(N),1}, \overline{X}_k^{(N),2}, \overline{R}_k^{(N),1}, \overline{R}_k^{(N),2}, \overline{I}_k^{(N)}, \{\overline{D}_{kl}^{(N),1}, \overline{D}_{kl}^{(N),2}, l \in \mathcal{K} \cup \{0\}\}, \\
&\overline{\nu}_0^{(N),1,k}, \overline{\nu}^{(N),1,k}, \overline{\nu}_0^{(N),2,k}, \overline{\nu}^{(N),2,k}, \overline{\eta}_0^{(N),1,k}, \overline{\eta}^{(N),1,k}, \overline{\eta}_0^{(N),2,k}, \overline{\eta}^{(N),2,k}, \\
&\overline{A}_{\cdot,\nu}^{(N),1,k}, \overline{D}_\cdot^{(N),1,k}, \overline{A}_{\cdot,\nu}^{(N),2,k}, \overline{D}_\cdot^{(N),2,k}, \overline{A}_{\cdot,\eta}^{(N),1,k}, \overline{S}_\cdot^{(N),1,k}, \overline{A}_{\cdot,\eta}^{(N),2,k}, \overline{S}_\cdot^{(N),2,k}\Big).
\end{aligned} \tag{5.47}$$

Then for each $k \in \mathcal{K}$, $\overline{Z}_k^{(N)}$ is a $\mathcal{Y}_k$-valued process, where $\mathcal{Y}_k$ is the space

$$\begin{aligned}
\mathcal{Y}_k \doteq\ &(\mathbb{R}_+)^2 \times (\mathcal{D}_{\mathbb{R}_+}[0, \infty))^{2K+8} \times \mathcal{M}_F[0, H_{1,k}^s) \times \mathcal{D}_{\mathcal{M}_F[0, H_{1,k}^s)}[0, \infty) \\
&\times \mathcal{M}_F[0, H_{2,k}^s) \times \mathcal{D}_{\mathcal{M}_F[0, H_{2,k}^s)}[0, \infty) \times \mathcal{M}_F[0, H_{1,k}^r) \times \mathcal{D}_{\mathcal{M}_F[0, H_{1,k}^r)}[0, \infty) \\
&\times \mathcal{M}_F[0, H_{2,k}^r) \times \mathcal{D}_{\mathcal{M}_F[0, H_{2,k}^r)}[0, \infty) \times (\mathcal{D}_{\mathcal{M}_F([0, H_{1,k}^s) \times \mathbb{R}_+)}[0, \infty))^2 \times (\mathcal{D}_{\mathcal{M}_F([0, H_{2,k}^s) \times \mathbb{R}_+)}[0, \infty))^2 \\
&\times (\mathcal{D}_{\mathcal{M}_F([0, H_{1,k}^r) \times \mathbb{R}_+)}[0, \infty))^2 \times (\mathcal{D}_{\mathcal{M}_F([0, H_{2,k}^r) \times \mathbb{R}_+)}[0, \infty))^2
\end{aligned}$$

equipped with the product metric. Clearly, $\mathcal{Y}_k$ is a Polish space. Let

$$\overline{Z}^{(N)} = (\overline{Z}_k^{(N)}, k \in \mathcal{K}). \tag{5.48}$$

Theorem 5.2. *Suppose Assumption 5.1 is satisfied. Then the sequence $\{\overline{Z}^{(N)}\}$ defined in (5.48) is relatively compact in the Polish space $\Pi_{k \in \mathcal{K}} \mathcal{Y}_k$, and is therefore tight.*

Proof. We follow closely the steps in [16] by adapting to the network setting, so we will next sketch the main steps in the proof and give the pointers to the proofs in [16]. By applying Kurtz' criteria (see Theorem 3.8.6 of [9] for details) and a similar argument for Lemma 5.8(2) in [18] together with the bounds in (5.45) and (5.46), we can show that under Assumption 5.1, for each $k \in \mathcal{K}$, the sequences $\{\overline{X}_k^{(N),1}\}$, $\{\overline{X}_k^{(N),2}\}$, $\{\overline{I}_k^{(N)}\}$, $\{\overline{L}_k^{(N),1}\}$, $\{\overline{L}_k^{(N),2}\}$, $\{\overline{R}_k^{(N),1}\}$, $\{\overline{R}_k^{(N),2}\}$, $\{\langle \mathbf{1}, \overline{\nu}^{(N),1,k} \rangle\}$, $\{\langle \mathbf{1}, \overline{\nu}^{(N),2,k} \rangle\}$, $\{\langle \mathbf{1}, \overline{\eta}^{(N),1,k} \rangle\}$, $\{\langle \mathbf{1}, \overline{\eta}^{(N),2,k} \rangle\}$, the sequences $\{\overline{D}_\varphi^{(N),j,k}\}$, $\{\overline{A}_{\varphi,\nu}^{(N),j,k}\}$, for every $j = 1, 2$ and $\varphi \in \mathcal{C}_b([0, H_{j,k}^s) \times \mathbb{R}_+)$, and the sequences $\{\overline{S}_\varphi^{(N),j,k}\}$, $\{\overline{A}_{\varphi,\eta}^{(N),j,k}\}$, for every $j = 1, 2$ and $\varphi \in \mathcal{C}_b([0, H_{j,k}^r) \times \mathbb{R}_+)$, are relatively compact. By a similar argument of Lemma 6.4 of [16], together with the fact that

$$\mathbb{E}\left[X_k^{(N),1}(0) + X_k^{(N),2}(0) + E_k^{(N)}(t) + I_k^{(N)}\right] < \infty,$$

we can show that under Assumption 5.1, for every $f \in \mathcal{C}_c^1(\mathbb{R}_+)$ and $k \in \mathcal{K}$, the sequences $\{\langle f, \overline{\nu}^{(N),j,k}\rangle\}$ and $\{\langle f, \overline{\eta}^{(N),j,k}\rangle\}$, $j = 1, 2$, of $\mathcal{D}_{\mathbb{R}}[0,\infty)$-valued random variables are relatively compact. In addition, with the application of the Jakubowski's criteria (cf. Proposition 6.5 of [16]), in the same way as the proofs of Lemmas 6.6 and 6.7 of [16], we can show that under Assumption 5.1, for each $k \in \mathcal{K}$, the sequences $\{\overline{\nu}^{(N),j,k}\}$ and $\{\overline{\eta}^{(N),j,k}\}$, $j = 1, 2$, are relatively compact and the sequences $\{\overline{D}_{\cdot}^{(N),j,k}\}$ and $\{\overline{A}_{\cdot,\nu}^{(N),j,k}\}$ are relatively compact in $\mathcal{D}_{\mathcal{M}_F([0,H_{j,k}^s)\times\mathbb{R}_+)}[0,\infty)$. Similarly, the sequences $\{\overline{S}_{\cdot}^{(N),j,k}\}$ and $\{\overline{A}_{\cdot,\eta}^{(N),j,k}\}$ are relatively compact in $\mathcal{D}_{\mathcal{M}_F([0,H_{j,k}^r)\times\mathbb{R}_+)}[0,\infty)$ for each $j = 1, 2$. The above results together with the direct application of Prohorov's theorem imply the tightness of the processes $\{\overline{Z}^{(N)}\}$. ∎

The rest of this section is devoted to proving Theorem 5.5 under Assumptions 5.1–5.2 that every limit of any subsequence of $\{\overline{Z}^{(N)}\}$ solves the fluid equations. This and the uniqueness of such a solution under Assumptions 3.1–3.3 in Theorem 3.3 together show that Theorem 5.1 holds. We note that Theorem 5.2 only requires Assumption 5.1 while Theorem 5.5 requires both Assumptions 5.1–5.2. We start with two supporting lemmas.

Lemma 5.3. *For each $k \in \mathcal{K}$ and $l \in \mathcal{K}\cup\{0\}$, $\overline{D}_{kl}^{(N),1} - P_{kl}\overline{D}_{\mathbf{1}}^{(N),1,k} \Rightarrow 0$ and $\overline{D}_{kl}^{(N),2} - P_{kl}\overline{D}_{\mathbf{1}}^{(N),2,k} \Rightarrow 0$ as $N \to \infty$.*

Proof. Fix $k \in \mathcal{K}$ and $l \in \mathcal{K} \cup \{0\}$. It follows from (5.8) and (5.30) that

$$\begin{aligned}
&\overline{D}_{kl}^{(N),1}(t) - P_{kl}\overline{D}_{\mathbf{1}}^{(N),1,k}(t) \\
= \;& \frac{1}{N}\sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(t)} \sum_{s\in[0,t]} (\mathbb{1}_{\{\phi^{1,k}(j)=e_l\}} - P_{kl})\mathbb{1}_{\left\{\frac{da_j^{(N),1,k}}{dt}(s-)>0,\ \frac{da_j^{(N),1,k}}{dt}(s+)=0\right\}}.
\end{aligned}$$

Since the service distributions $G_{1,k}^s$ hava a density, then with probability 1, any two external customers will not finish service at the same time, thus for each $T > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le T}(\overline{D}_{kl}^{(N),1}(t) - P_{kl}\overline{D}_{\mathbf{1}}^{(N),1,k}(t))^2\right] \\
\le\;& \frac{1}{N^2}\mathbb{E}\left[\sum_{j=-\mathcal{E}_k^{(N)}+1}^{E_k^{(N)}(T)} (\mathbb{1}_{\{\phi^{1,k}(j)=e_l\}} - P_{kl})^2\right] \\
\le\;& \frac{1}{N}\mathbb{E}\left[(\mathbb{1}_{\{\phi^{1,k}(1)=e_l\}} - P_{kl})^2\right]\mathbb{E}\left[\overline{E}_k^{(N)}(T) + \langle\mathbf{1}, \overline{\nu}_0^{(N),1,k}\rangle + \langle\mathbf{1}, \overline{\eta}_0^{(N),1,k}\rangle\right].
\end{aligned}$$

By Assumption 5.1,

$$\lim_{N\to\infty}\mathbb{E}\left[\sup_{0\le t\le T}(\overline{D}_{kl}^{(N),1}(t) - P_{kl}\overline{D}_{\mathbf{1}}^{(N),1,k}(t))^2\right] = 0$$

On the other hand, It follows from (5.9) and (5.31) that

$$\begin{aligned}
&\overline{D}_{kl}^{(N),2}(t) - P_{kl}\overline{D}_{\mathbf{1}}^{(N),2,k}(t) \\
= \;& \frac{1}{N}\sum_{j=-\mathcal{C}_k^{(N)}+1}^{I_k^{(N)}(t)} \sum_{s\in[0,t]} (\mathbb{1}_{\{\phi^{2,k}(j)=e_l\}} - P_{kl})\mathbb{1}_{\left\{\frac{da_j^{(N),2,k}}{dt}(s-)>0,\ \frac{da_j^{(N),2,k}}{dt}(s+)=0\right\}}.
\end{aligned}$$

Since the service distributions $G^s_{2,k}$ have a density, then with probability 1, any two internal customers will not finish service at the same time, thus for each $T > 0$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le T}(\overline{D}^{(N),2}_{kl}(t) - P_{kl}\overline{D}^{(N),2,k}_{\mathbf{1}}(t))^2\right] \\
\le\ & \frac{1}{N^2}\mathbb{E}\left[\sum_{j=-\mathcal{C}^{(N)}_k+1}^{I^{(N)}_k(t)} (\mathbb{1}_{\{\phi^{2,k}(j)=e_l\}} - P_{kl})^2\right] \\
\le\ & \frac{1}{N}\mathbb{E}\left[(\mathbb{1}_{\{\phi^{2,k}(1)=e_l\}} - P_{kl})^2\right]\mathbb{E}\left[\overline{I}^{(N)}_k(T) + \langle \mathbf{1}, \overline{\nu}^{(N),2,k}_0\rangle + \langle \mathbf{1}, \overline{\eta}^{(N),2,k}_0\rangle\right].
\end{aligned}$$

By Assumption 5.1 and (5.44),

$$\lim_{N\to\infty}\mathbb{E}\left[\sup_{0\le t\le T}(\overline{D}^{(N),2}_{kl}(t) - P_{kl}\overline{D}^{(N),2,k}_{\mathbf{1}}(t))^2\right] = 0$$

and hence the lemma is proved. ∎

Lemma 5.4. *For each $k \in \mathcal{K}$, $j = 1, 2$ and $T \in [0,\infty)$, as $N \to \infty$,*

$$\mathbb{E}\left[\sup_{t\in[0,T]}\left|\overline{A}^{(N),j,k}_{\theta^{(N)}_k,\eta}(t) - \int_0^t\left(\int_{[0,H^r_{j,k})}\mathbb{1}_{[0,\overline{\chi}_k(s)]}(u)h^r_{j,k}(u)\overline{\eta}^{j,k}_s(du)\right)ds\right|\right] \to 0. \tag{5.49}$$

Moreover, almost surely,

$$\overline{R}^j_k(t) = \int_0^t\left(\int_{[0,H^r_{j,k})}\mathbb{1}_{[0,\overline{\chi}_k(s)]}(u)h^r_{j,k}(u)\overline{\eta}^{j,k}_s(du)\right)ds,\ t\in[0,\infty), j = 1,2. \tag{5.50}$$

Proof. First, (5.50) follows directly from (5.49) and (5.43). Now, we focus on showing (5.49). Fix $k \in \mathcal{K}$, $j \in \{1,2\}$ and $T > 0$. It follows from the definition of $A^{(N),j,k}_{\theta^{(N)}_k,\eta}$ in (5.38) that for each $t \in [0,T]$,

$$\begin{aligned}
&\overline{A}^{(N,j,k)}_{\theta^{(N)}_k,\eta}(t) - \int_0^t\left(\int_{[0,H^r_{j,k})}\mathbb{1}_{[0,\overline{\chi}_k(s)]}(u)h^r_{j,k}(u)\overline{\eta}^{j,k}_s(du)\right)ds \\
=\ & \int_0^t\left(\int_{[0,H^r_{j,k})}\mathbb{1}_{[0,\chi^{(N)}_k(s-)]}(u)h^r_{j,k}(u)\,\overline{\eta}^{(N),j,k}_s(du)\right)ds - \int_0^t\left(\int_{[0,H^r_{j,k})}\mathbb{1}_{[0,\overline{\chi}_k(s)]}(u)h^r_{j,k}(u)\overline{\eta}^{j,k}_s(du)\right)ds \\
=\ & \int_0^t\left(\int_{[0,H^r_{j,k})}\left(\mathbb{1}_{[0,\chi^{(N)}_k(s-)]}(u) - \mathbb{1}_{[0,\overline{\chi}_k(s)]}(u)\right)h^r_{j,k}(u)\,\overline{\eta}^{(N),j,k}_s(du)\right)ds \\
+\ & \left[\int_0^t\left(\int_{[0,H^r_{j,k})}\mathbb{1}_{[0,\overline{\chi}_k(s)]}(u)h^r_{j,k}(u)\overline{\eta}^{(N),j,k}_s(du)\right)ds - \int_0^t\left(\int_{[0,H^r_{j,k})}\mathbb{1}_{[0,\overline{\chi}_k(s)]}(u)h^r_{j,k}(u)\overline{\eta}^{j,k}_s(du)\right)ds\right].
\end{aligned}$$

For each $t \in [0,T]$ and $\kappa \in [0, H^r_{j,k})$, let

$$\overline{C}^{(N),j}_1(t,\kappa) \doteq \left|\int_0^t\left(\int_{[0,H^r_{j,k})}\left(\mathbb{1}_{[0,\chi^{(N)}_k(s-)\wedge\kappa]}(u) - \mathbb{1}_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u)\right)h^r_{j,k}(u)\,\overline{\eta}^{(N),j,k}_s(du)\right)ds\right|, \tag{5.51}$$

$$\overline{C}^{(N),j}_2(t,\kappa) \doteq \left|\int_0^t\left(\int_{[0,H^r_{j,k})}\left(\mathbb{1}_{(\chi^{(N)}_k(s-)\wedge\kappa,\chi^{(N)}_k(s-)]}(u) - \mathbb{1}_{(\overline{\chi}_k(s)\wedge\kappa,\overline{\chi}_k(s)]}(u)\right)h^r_{j,k}(u)\,\overline{\eta}^{(N),j,k}_s(du)\right)ds\right|, \tag{5.52}$$

$$\overline{C}_3^{(N),j}(t,\kappa) \doteq \left| \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u) h^r_{j,k}(u) \overline{\eta}^{(N),j,k}_s(du) \right) ds \right. \\ \left. - \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u) h^r_{j,k}(u) \overline{\eta}^{j,k}_s(du) \right) ds \right| \tag{5.53}$$

and

$$\overline{C}_4^{(N),j}(t,\kappa) \doteq \left| \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{(\overline{\chi}_k(s)\wedge\kappa,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u) \overline{\eta}^{(N),j,k}_s(du) \right) ds \right. \\ \left. - \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{(\overline{\chi}_k(s)\wedge\kappa,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u) \overline{\eta}^{j,k}_s(du) \right) ds \right|. \tag{5.54}$$

Then, it is obvious that for each $t \in [0,T]$ and $\kappa \in [0, H^r_{j,k})$,

$$\left| \overline{A}^{(N,j,k)}_{\theta^{(N)}_k,\eta}(t) - \int_0^t \left(\int_{[0,H^r_{j,k})} \mathbb{1}_{[0,\overline{\chi}_k(s)]}(u) h^r_{j,k}(u) \overline{\eta}^{j,k}_s(du) \right) ds \right| \le \sum_{i=1}^4 \overline{C}_i^{(N),j}(t,\kappa).$$

From this, to prove (5.49), it suffices to show that for $i = 1,2,3,4$,

$$\lim_{\kappa \to H^r_{j,k}} \lim_{N\to\infty} \mathbb{E}\left[\sup_{0\le t\le T} \overline{C}_i^{(N),j}(t,\kappa) \right] = 0.$$

Since $h^r_{j,k}$ is locally bounded, let $\Xi^{j,k}_\kappa \doteq \sup_{0\le u\le\kappa} h^r_{j,k}$. It follows from Proposition 5.5 of [16] by taking $h = 1$ therein that for $s \ge 0$,

$$\eta^{(N),k}_s[0, \chi^{(N)}_k(s-)] = Q^{(N),1}_k(s) + Q^{(N),2}_k(s) + \iota^{(N)}_k(s),$$

where

$$\iota^{(N)}_k(s) \doteq \begin{cases} 0 & \text{if } (\chi^{(N)}_k(s-) - \chi^{(N)}_k(s))(L^{(N)}_k(s) - L^{(N)}_k(s-)) = 0, \\ 1 & \text{if } (\chi^{(N)}_k(s-) - \chi^{(N)}_k(s))(L^{(N)}_k(s) - L^{(N)}_k(s-)) > 0, \end{cases}$$

and $L^{(N)}_k(s) = L^{(N),1}_k(s) + L^{(N),2}_k(s)$. Firstly, note that $\chi^{(N)}_k(s-) \ge \chi^{(N)}_k(s)$ for all $s \ge 0$. It follows from (5.51) and (5.6) that

$$\begin{aligned} \sup_{0\le t\le T} \overline{C}^{(N),j}_1(t,\kappa) &\le \Xi^{j,k}_\kappa \int_0^T \left| \int_{[0,H^r_{j,k})} \left(\mathbb{1}_{[0,\chi^{(N)}_k(s-)\wedge\kappa]}(u) - \mathbb{1}_{[0,\overline{\chi}_k(s)\wedge\kappa]}(u) \right) \overline{\eta}^{(N),j,k}_s(du) \right| ds \\ &= \Xi^{j,k}_\kappa \int_0^T \left| \overline{\eta}^{(N),j,k}_s[0, \chi^{(N)}_k(s-)\wedge\kappa] - \overline{\eta}^{(N),j,k}_s[0, \overline{\chi}_k(s)\wedge\kappa] \right| ds \\ &\le \Xi^{j,k}_\kappa \int_0^T \left| \overline{\eta}^{(N),k}_s[0, \chi^{(N)}_k(s-)\wedge\kappa] - \overline{\eta}^{(N),k}_s[0, \overline{\chi}_k(s)\wedge\kappa] \right| ds \\ &\le \Xi^{j,k}_\kappa \int_0^T \left| \left(\overline{Q}^{(N),1}_k(s) + \overline{Q}^{(N),2}_k(s) + \overline{\iota}^{(N)}_k(s) \right) \wedge \overline{\eta}^{(N),k}_s[0,\kappa] - \overline{\eta}^{(N),k}_s[0, \overline{\chi}_k(s)\wedge\kappa] \right| ds, \end{aligned}$$

where $\overline{\iota}^{(N)}_k(s) = \iota^{(N)}_k(s)/N$. Since $\overline{E}_k$ and $\overline{\eta}^k_0 = \overline{\eta}^{1,k}_0 + \overline{\eta}^{2,k}_0$ are continuous, then by applying (4.3) of [16], $\overline{\eta}^k_s = \overline{\eta}^{1,k}_s + \overline{\eta}^{1,k}_s$ is also continuous. Thus, by the convergence of $\overline{Q}^{(N),1}_k$, $\overline{Q}^{(N),2}_k$ $\overline{\iota}^{(N)}_k$, $\overline{\eta}^{(N),1,k}$ and $\overline{\eta}^{(N),2,k}$ to $\overline{Q}^1_k$, $\overline{Q}^2_k$, $\mathbf{0}$, $\overline{\eta}^{1,k}$ and $\overline{\eta}^{2,k}$ respectively, we have for each $s \ge 0$,

$$\lim_{N\to\infty} \left(\left(\overline{Q}^{(N),1}_k(s) + \overline{Q}^{(N),2}_k(s) + \overline{\iota}^{(N)}_k(s) \right) \wedge \overline{\eta}^{(N),k}_s[0,\kappa] - \overline{\eta}^{(N),k}_s[0, \overline{\chi}_k(s)\wedge\kappa] \right) = 0.$$

Note that by (5.18) and (5.19),

$$\mathbb{E}\left[\int_0^T \left|\left(\overline{Q}_k^{(N),1}(s)+\overline{Q}_k^{(N),2}(s)+\overline{\iota}_k^{(N)}(s)\right)\wedge\overline{\eta}_s^{(N),k}[0,\kappa]-\overline{\eta}_s^{(N),k}[0,\overline{\chi}_k(s)\wedge\kappa]\right| ds\right]$$
$$\leq \mathbb{E}\left[\int_0^T\left(\overline{\eta}_s^{(N),k}[0,\kappa]+\overline{\eta}_s^{(N),k}[0,\kappa]\right)ds\right]\leq 2T\mathbb{E}\left[\langle \mathbf{1},\overline{\eta}_0^{(N),k}\rangle+\overline{E}_k^{(N)}(T)+\overline{I}_k^{(N)}(T)\right]<\infty.$$

This, together with an application of the dominated convergence theorem yields that

$$\lim_{\kappa\to H_{j,k}^r}\lim_{N\to\infty}\mathbb{E}\left[\sup_{0\leq t\leq T}\overline{C}_1^{(N),j}(t,\kappa)\right]=0.$$

Secondly, by (5.52) and an application of triangle inequality, we have that

$$\overline{C}_2^{(N),j}(t,\kappa)\leq 2\int_0^t\int_{[\kappa,H_{j,k}^r)}h_{j,k}^r(u)\,\overline{\eta}_s^{(N),j,k}(du)ds.$$

Moreover, by (5.54), we also have

$$\overline{C}_4^{(N),j}(t,\kappa)\leq\int_0^t\int_{[\kappa,H_{j,k}^r)}h_{j,k}^r(u)\overline{\eta}_s^{(N),j,k}(du)ds+\int_0^t\int_{[\kappa,H_{j,k}^r)}h_{j,k}^r(u)\overline{\eta}_s^{j,k}(du)ds.$$

Thus, by a similar argument in showing (7.30) of [16], we have

$$\lim_{\kappa\to H_{j,k}^r}\lim_{N\to\infty}\mathbb{E}\left[\sup_{0\leq t\leq T}\overline{C}_2^{(N),j}(t,\kappa)\right]=0 \text{ and } \lim_{\kappa\to H_{j,k}^r}\lim_{N\to\infty}\mathbb{E}\left[\sup_{0\leq t\leq T}\overline{C}_4^{(N),j}(t,\kappa)\right]=0.$$

Lastly, by using a similar argument as in Lemma 7.6 of [16], we have that

$$\lim_{\kappa\to H_{j,k}^r}\lim_{N\to\infty}\mathbb{E}\left[\sup_{0\leq t\leq T}\overline{C}_3^{(N),j}(t,\kappa)\right]=0. \qquad \blacksquare \tag{5.55}$$

Theorem 5.5. *Suppose that Assumptions 5.1–5.2 hold. Let $\overline{Z}$ be a weak limit of a subsequence of $\{\overline{Z}^{(N)}\}$, where for each $k\in\mathcal{K}$,*

$$\overline{Z}_k \doteq \Big(\overline{X}_k^1(0),\overline{X}_k^2(0),\overline{E}_k,\overline{X}_k^1,\overline{X}_k^2,\overline{R}_k^1,\overline{R}_k^2,\overline{I}_k,\{P_{kl}\overline{A}_{\mathbf{1},\nu}^{1,k},p_{kl}\overline{A}_{\mathbf{1},\nu}^{2,k},l\in\mathcal{K}\cup\{0\}\},\overline{\nu}_0^{1,k},\overline{\nu}^{1,k},\overline{\nu}_0^{2,k},\overline{\nu}^{2,k},$$
$$\overline{\eta}_0^{1,k},\overline{\eta}^{1,k},\overline{\eta}_0^{2,k},\overline{\eta}^{2,k},\overline{A}_{\cdot,\nu}^{1,k},\overline{A}_{\cdot,\nu}^{1,k},\overline{A}_{\cdot,\nu}^{2,k},\overline{A}_{\cdot,\nu}^{2,k},\overline{A}_{\cdot,\eta}^{1,k},\overline{A}_{\cdot,\eta}^{1,k},\overline{A}_{\cdot,\eta}^{2,k},\overline{A}_{\cdot,\eta}^{2,k}\Big)\in\mathcal{Y}_k.$$

Then $(\overline{X}^1,\overline{X}^2,\overline{\nu}^1,\overline{\nu}^2,\overline{\eta}^1,\overline{\eta}^2)$ solves the fluid equations.

Remark 5.6. *Due to Assumption 5.1, Theorem 5.2, Lemma 5.3 and the limits $\overline{M}_{\cdot,\nu}^{(N),j,k}=\overline{D}_{\cdot}^{(N),j,k}-\overline{A}_{\cdot,\nu}^{(N),j,k}\Rightarrow 0$ and $\overline{M}_{\cdot,\eta}^{(N),j,k}=\overline{S}_{\cdot}^{(N),j,k}-\overline{A}_{\cdot,\eta}^{(N),j,k}\Rightarrow 0$, $j=1,2$, from (5.41), any weak limit of any subsequence of $\{\overline{Z}^{(N)}\}$ has to take the form as $\overline{Z}$ in the statement of Theorem 5.5.*

Proof of Theorem 5.5. Denoting this subsequence again by $\overline{Z}^{(N)}$ and invoking the Skorokhod Representation Theorem, with a slight abuse of notation, we can assume that, $\mathbb{P}$ a.s., $\overline{Z}^{(N)}\to\overline{Z}$ as $N\to\infty$. Without loss of generality, we may further assume that the above convergence holds everywhere. Let $Y^{(N)}=(E^{(N)},X^{(N),1},X^{(N),2},\nu^{(N),1},\nu^{(N),2},\eta^{(N),1},\eta^{(N),2})$. Since $\overline{Z}^{(N)}\to\overline{Z}$ as $N\to\infty$, then it follows that, as $N\to\infty$,

$$(\overline{Y}_k^{(N)},\{\overline{D}_{kl}^{(N),1},\overline{D}_{kl}^{(N),2},l\in\mathcal{K}\cup\{0\}\})\to(\overline{Y}_k,\{P_{kl}\overline{A}_{\mathbf{1},\nu}^{1,k},p_{kl}\overline{A}_{\mathbf{1},\nu}^{2,k},l\in\mathcal{K}\cup\{0\}\}),$$

where $\overline{Y} = (\overline{E}, \overline{X}^1, \overline{X}^2, \overline{\nu}^1, \overline{\nu}^2, \overline{\eta}^1, \overline{\eta}^2)$. Together with (5.16), (5.17) and the fact that $\sum_{l\in\mathcal{K}\cup\{0\}} P_{kl} = 1$, this implies that

$$\overline{X}^1_k = \overline{X}^1_k(0) + \overline{E}_k - \overline{R}^1_k - \overline{A}^{1,k}_{\mathbf{1},\nu} \text{ and } \overline{X}^2_k = \overline{X}^2_k(0) + \overline{I}_k - \overline{R}^2_k - \overline{A}^{2,k}_{\mathbf{1},\nu}. \tag{5.56}$$

Moreover, by the same argument in getting (7.2) of [16], we have that for $j = 1, 2$,

$$\overline{A}^{j,k}_{\varphi,\nu} = \int_0^{\cdot} \langle \psi(\cdot, s) h^s_{j,k}(\cdot, s), \overline{\nu}^{j,k}_s \rangle ds. \tag{5.57}$$

On substituting (5.57) into (5.56), we see that the fluid equations (2.17) and (2.18) are satisfied. Next, Lemma 5.4 establishes the representations (2.15) and (2.16) in the fluid equations.

Fix $t \in [0, \infty)$ and $j = 1, 2$ such that for each $k \in \mathcal{K}$, $\overline{\nu}^{(N),j,k}_t \xrightarrow{w} \overline{\nu}^{j,k}_t$, $\overline{\eta}^{(N),j,k}_t \xrightarrow{w} \overline{\eta}^{j,k}_t$, $\overline{E}^{(N)}_k(t) \to \overline{E}_k(t)$, $\overline{X}^{(N),j}_k(t) \to \overline{X}^j_k(t)$, $\overline{R}^{(N),j}_k(t) \to \overline{R}^j_k(t)$, $\overline{I}^{(N)}_k(t) \to \overline{I}_k(t)$, $\overline{A}^{(N),j,k}_{\cdot,\nu}(t) \xrightarrow{w} \overline{A}^{j,k}_{\cdot,\nu}(t)$, $\overline{D}^{(N),j,k}_{\cdot}(t) \xrightarrow{w} \overline{A}^{j,k}_{\cdot,\nu}(t)$, $\overline{A}^{(N),j,k}_{\cdot,\eta}(t) \xrightarrow{w} \overline{A}^{j,k}_{\cdot,\eta}(t)$, $\overline{S}^{(N),j,k}_{\cdot}(t) \xrightarrow{w} \overline{A}^{j,k}_{\cdot,\eta}(t)$ as $N \to \infty$. Since $\overline{Z}^{(N)} \to \overline{Z}$ a.s., this occurs for t outside a countable set. By (5.57), this implies that for each $k \in \mathcal{K}$, $j = 1, 2$, as $N \to \infty$,

$$\overline{D}^{(N),j,k}_{\varphi}(t) \to \overline{A}^{j,k}_{\varphi,\nu}(t) = \int_0^t \langle \varphi(\cdot, s) h^s_{j,k}(\cdot, s), \overline{\nu}^{j,k}_s \rangle \, ds, \qquad \varphi \in \mathcal{C}_b([0, H^s) \times \mathbb{R}_+). \tag{5.58}$$

An analogous argument also implies that for each $k \in \mathcal{K}$ and $j = 1, 2$, as $N \to \infty$,

$$\overline{S}^{(N),j,k}_{\psi}(t) \to \overline{A}^{j,k}_{\psi,\eta}(t) = \int_0^t \langle \psi(\cdot, s) h^r_{j,k}(\cdot, s), \overline{\eta}^{j,k}_s \rangle \, ds, \qquad \psi \in \mathcal{C}_b([0, H^r) \times \mathbb{R}_+).$$

In particular, when $\varphi = \psi = \mathbf{1}$, the above two displays imply that (2.5) holds. Also, we immediately obtain that for each $k \in \mathcal{K}$ and $j = 1, 2$, as $N \to \infty$, $\langle \mathbf{1}, \overline{\nu}^{(N),j,k}_t \rangle \to \langle \mathbf{1}, \overline{\nu}^{j,k}_t \rangle$ and $\langle \mathbf{1}, \overline{\eta}^{(N),j,k}_t \rangle \to \langle \mathbf{1}, \overline{\eta}^{j,k}_t \rangle$. When combining with (5.22), (5.20), (5.16), (5.6),(5.23), (5.21), (5.17), (5.50), (5.10), Lemma 5.3 and the non-idling condition, this implies that all the equations in Definition 2.1 are satisfied at time t except (2.6), (2.8), (2.10) and (2.11).

It only remains to show that (2.6), (2.8), (2.10) and (2.11) are also satisfied at time t. We shall just prove (2.6). The same argument will also show that (2.8), (2.10) and (2.11) hold. By a similar argument as the proof of Theorem 2.1 in [16], in the fluid scale, we have

$$\begin{aligned} \left\langle \varphi(\cdot, t), \overline{\nu}^{(N),1,k}_t \right\rangle &= \left\langle \varphi(\cdot, 0), \overline{\nu}^{(N),1,k}_0 \right\rangle + \int_0^t \left\langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}^{(N),1,k}_s \right\rangle \, ds \\ &\quad - \overline{D}^{(N),1,k}_{\varphi}(t) + \int_{[0,t]} \varphi(0, s) d\overline{L}^{(N),1}_k(s). \end{aligned}$$

Since $\overline{\nu}^{(N),1,k}_0 \xrightarrow{w} \overline{\nu}^{1,k}_0$ by Assumption 5.1(iii), $\overline{\nu}^{(N),1,k}_s \xrightarrow{w} \overline{\nu}^{1,k}_s$ for a.e. $s \in [0, t]$, $\overline{\nu}^{(N),1,k}_t \xrightarrow{w} \overline{\nu}^{1,k}_t$ by our choice of t, and functions $\varphi(\cdot, t)$ and $\varphi_x(\cdot, s) + \varphi_s(\cdot, s)$, $s \in [0, t]$, are bounded and continuous, as $N \to \infty$, we have

$$\left\langle \varphi(\cdot, t), \overline{\nu}^{(N),1,k}_t \right\rangle \to \left\langle \varphi(\cdot, t), \overline{\nu}^{1,k}_t \right\rangle \quad \text{and} \quad \left\langle \varphi(\cdot, 0), \overline{\nu}^{(N),1,k}_0 \right\rangle \to \left\langle \varphi(\cdot, 0), \overline{\nu}^{1,k}_0 \right\rangle,$$

and, by the bounded convergence theorem,

$$\int_0^t \left\langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}^{(N),1,k}_s \right\rangle \, ds \to \int_0^t \left\langle \varphi_x(\cdot, s) + \varphi_s(\cdot, s), \overline{\nu}^{1,k}_s \right\rangle \, ds.$$

On the other hand, using an integration-by-parts argument, the facts that $\overline{L}^{(N),1}_k(0) = 0$, $\overline{L}^{(N),1}_k \to \overline{L}^1_k$, $\overline{L}^1_k$ is non-decreasing and $\varphi_s(0, \cdot)$ is bounded and continuous on $[0, t]$, along with the bounded convergence theorem, we see that, as $N \to \infty$, $\int_{[0,t]} \varphi(0, s) d\overline{L}^{(N),1}_k(s) \to \int_{[0,t]} \varphi(0, s) d\overline{L}^1_k(s)$. Combining the last four displays with (5.58), it follows that (2.6) holds. Then it follows that all fluid equations are satisfied for all but countably many t. By right-continuity (with respect to t) of each

of the terms in all fluid equations, we conclude that all fluid equations are a.s. satisfied for all $t \in [0, \infty)$. This completes the proof of the desired result that $(\overline{X}^1, \overline{X}^2, \overline{\nu}, \overline{\eta}^1, \overline{\eta}^2)$ satisfies the fluid equations. ∎

References

[1] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov. Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective. *Stochastic Systems.* Vol. 5 No. 1, 146–194. 2015.

[2] R. Atar, W. Kang, H. Kaspi and K. Ramanan. Long-time limit of nonlinearly coupled measure-valued equations that model many-server queues with reneging. *SIAM Journal on Mathematical Analysis.* Vol. 55, No. 6, 7189-7239, 2023.

[3] R. Atar, H. Kaspi and N. Shimkin. Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research.* Vol. 39, No. 3, 672–696. 2014.

[4] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing science perspective. *Journal of the American Statistical Association.* Vol. 100, No. 469, 36–50. 2005.

[5] J.G. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research.* Vol. 35, 347–362, 2010.

[6] J.G. Dai, S. He and T. Tezcan. Many-server diffusion limits for $G/Ph/n + GI$ queues. *Annals of Applied Probability.* Vol. 20, 1854–1890, 2010.

[7] J. G. Dai and S. He. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering.* Vol. 21, No. 1, 1–36, 2012.

[8] D. Gamarnik and D.A. Goldberg. Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. *Annals of Applied Probability.* Vol. 23, No. 6, 2382–2419, 2012.

[9] S.N. Ethier and T.G. Kurtz, *Markov processes: Characterization and convergence*, Wiley, 1986.

[10] N. Gans, G. Koole and A. Mandelbaum. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service and Operations Management* Vol. 5, 79–141, 2003.

[11] O. Garnett, A. Mandelbaum and M.I. Reiman. Designing a call center with impatient customers. *Manufacturing Service and Operations Management* Vol. 4, 208–227, 2002.

[12] R.W. Hall. *Patient Flow: Reducing Delay in Healthcare.* Springer, 2010.

[13] W. Kang. Fluid limits of many-server retrial queues with nonpersistent customers. *Queueing Systems.* Vol. 79, 183–219, 2015.

[14] W. Kang. Well-posedness and Sensitivity Analysis of a Fluid Model for Multiclass Many-Server Queues with Abandonment Under Global FCFS Discipline. *Revision.* 2024.

[15] W. Kang and G. Pang. Equivalence of fluid models for $G_t/GI/N + GI$ queues. In: Yin, G., Zhang, Q. (eds) Modeling, Stochastic Control, Optimization, and Applications. The IMA Volumes in Mathematics and its Applications, Vol 164. Springer. 2019.

[16] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Annals of Applied Probability.* Vol. 20, No. 6, 2204–2260, 2010.

[17] W. Kang and K. Ramanan. Asymptotic approximations for the stationary distributions of many-server queues with abandonment. *Annals of Applied Probability.* Vol. 22, No. 2, 477–521, 2012.

[18] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Annals of Applied Probability.* Vol. 21, No. 1, 33-114, 2011.

[19] H. Kaspi and K. Ramanan. SPDE limits for many-server queues. *Annals of Applied Probability.* Vol. 23, No. 1, 145–229, 2013.

[20] P. Khudyakov. *Statistical Analysis of Call Center data.* Ph.D. Thesis, Technion, July, 2010.

[21] G. Koole. *Call Center Optimization.* First Edition. MG Books, Amsterdam. 2013.

[22] J.S.H. van Leeuwaarden, B.W.J. Mathijsen and B. Zwart. Economies-of-Scale in many-server queueing systems: tutorial and partial review of the QED Halfin–Whitt heavy-traffic regime. *SIAM Review* Vol. 61, No. 3, 403–440, 2019.

[23] Y. Liu and W. Whitt. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems.* Vol. 71, No. 4, 405–444, 2012.

[24] Y. Liu and W. Whitt. A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Operations Research Letters.* Vol. 40, No. 5, 307–312, 2012.

[25] Y. Liu and W. Whitt. Large-time asymptotics for the $G_t/M_t/s_t+GI_t$ many-server fluid queue with abandonment. *Queueing Systems.* Vol. 71, No. 4, 405–444. 2012.

[26] Y. Liu and W. Whitt. Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability.* Vol. 24, No. 1, 378–421, 2014.

[27] Y. Liu and W. Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*. Vol. 59, 835–846, 2011.

[28] Y. Liu and W. Whitt. Algorithms for time-varying networks of many-server fluid queues. *INFORMS Journal on Computing*. Vol. 26, No. 1, 59–73, 2014.

[29] Z. Long and J. Zhang. A note on many-server fluid models with time-varying arrivals. *Probability in the Engineering and Informational Sciences.* Vol. 33, No. 3, 417–437, 2019.

[30] Z. Long and J. Zhang. Convergence to equilibrium states for fluid models of many-server queues with abandonment. *Operations Research Letters*. Vol. 42, No. 6-7, 388–393, 2014.

[31] Z. Long, N. Shimkin, H. Zhang and J. Zhang. Dynamic scheduling of multiclass many-server queues with abandonment: The generalized $c\mu/h$ rule. *Operations Research.* Vol. 68, No. 4, 1218–1230, 2020.

[32] Z. Long, T. Tezcan and J. Zhang. Routing and staffing in customer service chat systems with generally distributed service and patience times. *Manufacturing and Service Operations Management*. Vol. 26, No. 5, 1674–1691, 2024.

[33] A. Mandelbaum, W. Massey and M. Reiman, Strong approximations for Markovian service networks, *Queueing Systems*. Vol. 30, 149–201, 1998.

[34] A. Mandelbaum and P. Momcilovic, Queues with many servers and impatient customers. *Mathematics of Operations Research*. Vol. 37, No. 1, 41–65, 2012.

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

[36] A. Mandelbaum and S. Zeltyn, Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. *Technical Report*, Technion Institute of Technology, Israel, 2005.

[37] A. Mandelbaum and S. Zeltyn. Data stories about (im)patient customers in tele-queues. *Queueing Systems.* Vol. 75, 115–146, 2013.

[38] G. Pang, R. Talreja and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys.*, Vol. 4, 193–267, 2007.

[39] K.R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967.

[40] 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.* Vol. 47, No. 2, 1192–1228, 2022.

[41] J. Reed. The G/GI/N queue in the Halfin-Whitt regime. *Annals of Applied Probability.* Vol. 19, No. 6, 2211–2269. 2009.

[42] J. Reed and Y. Y. Shaki. A fair policy for the $G/GI/N$ queue with multiple server pools. *Mathematics of Operations Research*. Vol. 40, No. 3, 2014.

[43] P. Shi, M. Chou, J.G. Dai, D. Ding, and J. Sim. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Science.* Vol. 62, No. 1, 1–28, 2015.

[44] A. Walsh-Zuniga. Fluid limits of many-server queues with abandonment, general service time and continuous patience distributions. *Stochastic Processes and their Applications.* Vol. 124, No. 3, 1436–1468, 2014.

[45] A. R. Ward. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science*. Vol. 17, No. 1, 1–14, 2012.

[46] W. Whitt. Fluid models for multiserver queues with abandonments. *Operations Research.* Vol. 54, No. 1, 37–54, 2006.

[47] W. Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, 2002.

[48] G. Yom-Tov. *Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime.* Ph.D. Thesis, Technion, June, 2010.

[49] J. Zhang. Fluid models of many-server queues with abandonment. *Queueing Systems.* Vol. 73, 147–193, 2013.